Métodos de aprendizaje de máquina para el análisis automático de hipernasalidad en niños con labio y paladar hendido

Elizabeth Londoño Mora

Ingeniera Electrónica

Tutor
Prof.Dr.-Ing. Juan Rafael Orozco-Arroyave

Universidad de Antioquia
Facultad de ingeniería
Pregrado Ingeniería Electrónica
Medellín
2022

Cita	(Londoño Mora, E., (2022))
Referencia **Estilo APA 7 (2020)**	Londoño Mora, E., (2022). *Métodos de aprendizaje de máquina para el análisis automático de hipernasalidad en niños con labio y paladar hendido* [Modalidad presencial]. Universidad de Antioquia, Medellín

Grupo de Investigación en Telecomunicaciones Aplicadas (GITA)

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Césped es

Decano/director: Jesús Francisco Vargas Bonilla

Jefe departamento: Augusto Enrique Salazar Jiménez.

1- RESUMEN

Los niños que han sido operados de labio y paladar hendido pueden presentar algunos problemas del habla. Uno de los problemas más comunes es la nasalización excesiva, la cual afecta la producción de ciertos sonidos y reduce la inteligibilidad del habla.

En este trabajo, se presentan algunas técnicas de aprendizaje de y procesamiento de señales adecuadas para la evaluación automática del nivel de nasalización en la voz de niños que han sido operados de labio y paladar hendido.

Para esto, se consideraron 88 grabaciones de niños cuyo lenguaje nativo es el español de Colombia y en inglés se tiene 44 grabaciones de niños cuyo lenguaje nativo es el inglés Hindú.

Los pacientes de las grabaciones en español leyeron 6 frases, y fueron evaluados por un especialista, el cual los asigno en uno de cuatro posibles niveles de nasalización: 0, 1, 2 y 3 (donde 0 son personas sanas mientras que tres son personas con el máximo nivel de nasalización), mientras que los pacientes de las grabaciones en inglés leyeron 3 frases y para este caso los pacientes fueron divididos en dos grupos los cuales son las grabaciones de control (personas sanas) y personas con hipernasalización.

El trabajo se divide en dos partes importantes las cuales son: primero la identificación de nasalización, donde en este momento lo que se realiza es aplicar a las diferentes grabaciones, una red neuronal pre-entrenada con diferentes métodos como lo son no supervisado, es decir analizar todos los fonemas postriors así no se encuentren presentes en la grabación, semi-supervisado, analizar solo los fonemas posterios presentes en el audio y supervisado para este se utiliza los que es el alineamiento forzado y de esta manera analizar cada fonema en el momento de tiempo en que son pronunciados, y con la comparación de estos tres métodos se busca identificar si es posible cuantificar la precisión con la que se producen algunos sonidos nasales, en un segundo momento se aplican diferentes métodos de clasificación como lo son máquina de soporte vectorial multiclase y PCA (este método solo es aplicado al mejor resultado obtenido de svm) a los resultados obtenidos por medio del alineamiento forzado, donde por medio de estos métodos de clasificación es posible detectar el nivel de nasalización con un acierto máximo del 56.0 %.

LISTA DE FIGURAS

LISTA DE TABLAS

CONTENIDO

2 INTRODUCCIÓN

El labio y paladar hendido (CLP, del inglés cleft lip and palate) es una de las malformaciones congénitas más frecuentes en todo el mundo, representa el 2-3 de estas, y existen varias patologías del habla asociadas a ella, tales como: hipernasalidad, oclusión glotal, entre otros.

Hipernasalidad es la patología más común presente en pacientes con CLP, esta es la incapacidad para lograr un cierre velofaríngeo adecuado, durante el habla que da como resultado la percepción de hipernasalidad, caracterizada por resonancia nasal anormal debido al paso excesivo de aire a través cavidad nasal.

La percepción de la hipernasalidad es una tarea compleja debido a que los fonoaudiólogos deben confiar en su capacidad auditiva (es decir depende de la capacidad experiencia de los especialistas), una persona a lo largo de su evaluación puede presentar diferentes diagnósticos, también su diagnóstico se puede tornar complicado debido a que un especialista debe recibir pacientes a lo largo del día, por lo que al final de una jornada sus sentidos pueden encontrarse disminuidos.

Por esta razón es que se hace necesario el estudio e implementación de diferentes meto- dos de aprendizaje, para la detección de hipernasalidad, y posterior migración a otras patologías, relacionadas con el habla.

Teniendo en cuenta los aspectos antes mencionados en este trabajo, se propone método de aprendizaje que ayude a evaluar los niveles de hipernasalidad.

En este trabajo se propone analizar el nivel de hipernasalidad en niños con labio de paladar hendido utilizando métodos de aprendizaje de máquina para cuantificar la calidad de producción de ciertos sonidos.

3 OBJETIVOS

3.1.1. OBJETIVO GENERAL

Implementar diferentes métodos de aprendizaje profundo (como lo son métodos no supervisados y métodos semi-supervisado y supervisados) como las redes neuronales recurrentes para cuantificar la precisión con la que se producen los sonidos que se ven afectados por algunos de los trastornos del habla derivados del labio y paladar hendido.

3.1.2. OBJETIVOS ESPECÍFICOS

- Identificar las características que son útiles para detectar el nivel de nasalización en la voz.
- Aplicar métodos de reconocimiento automático de voz para transcribir y analizar las grabaciones.
- Implementar métodos de aprendizaje profundo que sean capaces de cuantificar el nivel de nasalización.
- Validar el sistema implementado por medio de pruebas con diferentes muestras (diferentes idiomas español e inglés) de tal modo que quede comprobado que el sistema no está sesgado.

4. RESULTADOS ESPERADOS

Los resultados que se esperan al momento de culminar la etapa práctica en el grupo de investigación es que aplicando los diferentes métodos de caracterización como lo son alineamiento forzado, mapas fonémicos, una red neuronal sobre las diferentes oraciones se podrá validar las hipótesis iniciales, la cual era que posible detectar el nivel de nasalización a partir de señales de voz, utilizando métodos de aprendizaje automático, combinado con técnicas de procesamiento de señales.

Una vez validada esta hipótesis hará entrega de un manual en forma de manuscrito que describe los métodos y el paso a seguir para obtener los resultados esperados con cualquier tipo de grabación, del mismo modo los códigos utilizados se encontrarán en un repositorio donde podrán ser descargados libremente.

5. TRABAJOS RELACIONADOS

Existen trabajos que han considerado análisis de aprendizaje de máquina para análisis de hipernasalidad a partir de señales de voz, los cuales se describen algunos a continuación:

S. Murillo Rendon. et al en el año 2011. Realizaron el análisis de 110 voces sanas y 156 grabaciones de pacientes con CLP, para esto consideran las cinco vocales españolas. Las características que extrajeron fueron 11 coeficientes Cepstrales en las Frecuencias de Mel (MFCC) y el análisis automático lo realizaron por medio de máquina de soporte vectorial (SVM), donde se puede evidenciar un aumento del 20 % cuando se consideran los Coeficientes ceptrales. J.R. Orozco-Arroyave, et al en el año 2011. Realizaron el análisis de 110 voces sanas y 156 grabaciones de pacientes con CLP, para esto consideran las cinco vocales españolas las características que extrajeron fueron MFCC, análisis de componentes principales (PCA) y el análisis auto- mágico lo realizaron por medio de SVM, Los autores reportaron aciertos de hasta 20 % en la detección automática de hipernasalidad. Ling He. et al en el año 2014. Realizaron el análisis de 567 pacientes que tienen una reparación primaria del paladar hendido, para esto consideran 63 palabras de uso común que cubren las 21 consonantes iniciales y la más amplia usa vocales en mandarín. Las características que se extrajeron fueron MFCC y el análisis automático lo realizaron por medio de Red neuronal convolucional (del inglés Convolutional neural network CNN), La precisión de clasificación para cuatro niveles de hipernasalidad. Alcanza hasta el 83 % y la identificación correcta de la omisión de consonantes supera el 94 %. Ling He. et al en el año 2015. Realizaron el análisis 567 grabaciones de niños (de 5 a 12 años de edad), las grabaciones se recogen de 30 niños (15 varones y 15 mujeres), para esto consideran 63 palabras de uso común que cubren las 21 consonantes iniciales y la más amplia usa vocales en mandarín. Las características que se extrajeron fueron MFCC y el análisis automático lo realizaron por medio del algoritmo K-Nearest-Neighbor (KNN), se halló que las precisiones de clasificación en conjuntos de datos separados por género son una un poco más alto que en el conjunto de datos de género mixto. Marzieh Golabbakhsh, et al en el año 2017. Realizaron el análisis de 25 pacientes normales y 25 pacientes con CLP, para esto consideran seis oraciones diferentes las características que extrajeron fueron MFCC y el análisis automático lo realizaron por medio de CNN Se obtuvo un 85 % con una sensibilidad del 82 % y una especificidad del 85 %. Xiyue Wang. et al en el año 2015. Realizaron el análisis de 48 pacientes con hipernasalidad (24 hombres y 24 mujeres) y 48 controles (24

hombres y 24 mujeres), para esto consideran 167 sílabas sin repetición, las características que extrajeron fueron MFCC y el análisis automático lo realizaron por medio de DNN, puede lograr un rendimiento de detección, ya que considera más información de frecuencia. Vikram C, et al en el año 2020. Realizaron el análisis de 40 hombres,35 mujeres, para esto se consideran consonantes nasales (NC), consonantes orales (OC), vocales nasalizadas (NV) y vocales orales (OV), las características que extrajeron fueron MFCC y el análisis automático lo realizaron por medio de DNN, PCC de 0.689 para predecir la gravedad de la hipernasalidad de habla de adultos y 0.651 para casos de habla de niños. Henna Raunak Seth Tak, en el año 2016. Et al. realizaron el análisis de 30 niños operados para CLP, y 30 niños con desarrollo típico, para esto consideran oraciones orales especialmente construidas en marathi, una tarea de conteo de números, las características que se extrajeron fueron MFCC este análisis no es automatizado por lo que no se tienen en cuenta características ni análisis automático. Viviane Cristina de Castro Marino. et al en el año de 2020. Realizaron el análisis 80 individuos con CLP, para esto consideran nueve estímulos de habla diferentes, incluido el conteo y oraciones cortas caracterizadas por sonidos orales, este análisis no es automatizado por lo que no se tienen en cuenta características ni análisis automático. Vikram C. Mathad, et al en el año. 2021. realizaron el análisis 60 niños con parálisis cerebral, y 10 controles, subconjunto de 38 niños (28 PC y 10 controles), para esto consideran de cada niño, 24 oraciones con diferentes consonantes objetivo las características que extranjeros fueron 13 FCC y el análisis automático lo realizaron por medio de BTLS (Basic Trauma Life Support) se llegó a la conclusión que la nasalidad moderada se puede identificar tanto a partir de vocales altas como bajas, mientras que las vocales altas y bajas, así como las consonantes sonoras, se ven afectadas en casos hipernasales severos- Vikram Cmathad. Et al. 2021. Realizaron el análisis se realiza a una base de datos que contiene muestras de voz leída en ingles grabadas transcripciones ortográficas para cada frase leída, las características que extranjeros fueron 39 MFCC y el análisis automático lo realizaron por medio de SVM y de DNN, las bases de datos como para estos resultados empíricamente muestran que la medida objetiva de hipernasalidad (del inglés the objective hypernasality measure OHM) era robusto para una variedad de condiciones en las grabaciones de las oraciones.

En general los trabajos relacionados han considerado como análisis acústico los MFCC y como métodos de aprendizaje algunos con SVM y otros con métodos de aprendizaje profundo.

La contribución principal de este trabajo es que se agregara un análisis de cómo se afectan la precisión de articulación considerando análisis fonémico por medio de una red neuronal recurrente y métodos de procesamiento de señales.

6 MARCO TEÓRICO

En el tratamiento de niños con CLP corregido, se pueden presentar problemas de resonancia y emisión vocal, tales como: Hiponasalidad e Hipernasalidad. Se indica que es más frecuente encontrar casos con hipernasalidad (90 %), mientras la Hiponasalidad ocurre en un (10 %). El interés en la detección de la hipernasalidad está relacionado con que su presencia indica problemas anatómicos, neurológicos y del sistema nervioso periférico. La presencia de hipernasalidad, entendida como el escape de aire nasal y articulaciones compensatorias, conlleva a la baja inteligibilidad de la voz, la cual ocasiona un deterioro de la comunicación con su entorno que se manifiesta en cambios de actitud interpersonal y de comportamiento.

6.1 NASALIZACIÓN Y EMISIÓN NASAL

La nasalización se define como la comunicación que existe entre la cavidad nasal y el resto del tracto vocal; mientras que la emisión nasal se refiere al escape anormal de aire por la ruta nasal. Este escape anormal reduce la presión intraoral causando distorsión en las consonantes. Cuando el escape de aire resulta en un resoplo audible, la emisión nasal es más perceptible y el habla es seriamente afectada. La nasalidad, comúnmente llamada hipernasalidad, se refiere a la baja calidad de voz, que resulta de la adición inapropiada del sistema de resonancia nasal al tracto vocal.

En contraste a la emisión nasal, la nasalidad no envuelve grandes flujos de aire nasal, por lo que no hay cambios significativos en la presión de aire intraoral.

6.2 FONÉTICA CLÍNICA

Es ya conocido que el campo de la fonética ha estudiado de manera científica los sonidos del habla desde básicamente tres aspectos: producción y trascripción (fonética articulatoria), propiedades físicas (fonética acústica) y propiedades aerodinámicas (fonética aerodinámica). Por su parte, la fonética clínica, se ha ocupado de estos mismos aspectos, pero para el habla de personas con desórdenes o habla anómala. Los estudios fonéticos clínicos han mostrado tres tendencias según la forma como han abordado y analizado los datos. Así, se encuentran estudios que han descrito perceptualmente (a través de la audición) la articulación del habla, los que han utilizado instrumentación y los que han

combinado las dos, uso de instrumentación con descripción perceptual.

6.3 CARACTERIZACIÓN DE SEÑALES DE VOZ

El proceso de caracterización consiste en la extracción de medidas acústicas de una señal de voz por medio de las cuales se puedan identificar algunos problemas en el habla que afectan la voz en aspectos como la fonación, articulación y prosodia. Actualmente, no existe un conjunto estándar de medidas acústicas para la caracterización de señales de voz, sin embargo, existen diversas mediciones que reflejan alteraciones en la voz de personas con CLP.

6.3.1 EXTRACCIÓN DE CARACTERÍSTICAS

Dado que la precisión del sistema depende de los rasgos utilizados para caracterizar las señales de voz, y que no hay un conocimiento claro sobre qué características proporcionan una mejor caracterización de la hipernasal, es por ello que se llevan diferentes técnicas de selección las cuales son:

6.3.1.1 COEFICIENTES EN LA ESCALA DE MEL (MFCC)

Estos son un tipo particular de coeficientes cepstrales derivados de la aplicación del Cepstrum sobre una ventana de tiempo de la señal de voz. Analizando el Cepstrum desde un punto de vista matemático, podemos decir que se trata de un operador que transforma una convolución en el tiempo en una suma en el dominio espectral. De esta forma se consigue separar de una forma elegante las dos componentes de información de la señal de voz: la excitación y el tracto vocal. En general, el Cepstrum se define como la transformada inversa de Fourier del logaritmo del espectro de la señal de voz.

De aquí surge el concepto de coeficientes MFCC que hacen uso de una nueva escala de frecuencia no lineal denominada MEL para imitar el comportamiento psico acústico a tonos puros de distinta frecuencia dentro del oído humano. De hecho, estudios dentro de esta ciencia han demostrado que el sistema auditivo humano procesa la señal de voz en el dominio espectral, caracterizándose por tener mayores resoluciones en bajas frecuencias y esto es precisamente lo que se consigue mediante la escala MEL, asignar mayor relevancia

a las bajas frecuencias de forma análoga a como se hace en el sistema auditivo humano, en concreto en el oído interno.[2] Donde los pasos que se siguen para el cálculo de MEL son los siguientes:

- Calcular la transformada de Fourier en tiempo corto (espectrograma)
- Seleccionar el número de filtros (algunos de los valores típicos son 32, 64, 128) donde este parámetro ayuda a controlar la resolución en frecuencia del espectro de MEL.
- Elegir frecuencias de corte superior e inferior donde los valores típicos para estas frecuencias son de 50 Hz y 8 kHz.
- Construir el banco de filtros.
- Aplicar el banco de filtros a la transformada de Fourier en tiempo corto.
- Aplicar la transformada discreta del coseno.

Donde la ecuación para convertir de frecuencia lineal a MEL es la siguiente:

$$m = 1125\ln(1 + fHz/700)$$ *Ecuación (1)*

En la *Figura 1* se muestra el ejemplo de la transformada de Fourier en tiempo corto, un banco de filtros y el espectro de Mel resultante:

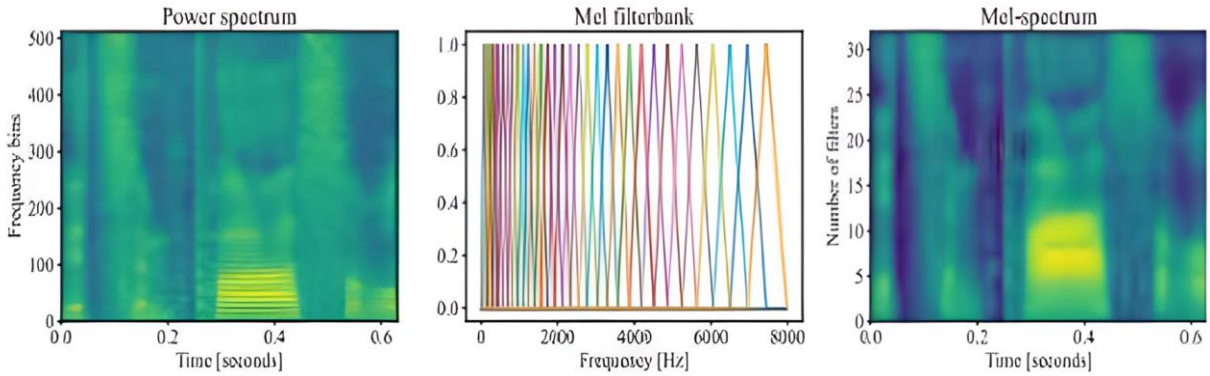

Figura 1*- Proceso del paso por los filtros de MEL*

6.3.1.2 EXTRACCIÓN DE FONEMAS

El método considerado en este trabajo para medir la precisión de articulación de fonemas consiste en convertir una secuencia de enventanado S→t = {→s0, →s1, ..., →sT−1} , en una secuencia de probabilidades de Y→t[→z] = {→y0[→z], →y1[→z], ..., →yT−1[→z]}, donde →z = 1, 2, . . ., z, . . ., Z son todos los grupos de fonemas posibles; por lo tanto, y[z] es la probabilidad de ocurrencia de la clase de fonema z-esima en el cuadro de voz t-esima.

Las probabilidades posteriores se calculan utilizando una red recurrente multietiqueta con capas convolucionales, que se puede utilizar para el reconocimiento automático de secuencias de fonemas en función de la probabilidad de ocurrencia del fonema. La precisión del fonema se evalúa considerando la forma de articulación, que se refiere a cómo se configuran los articuladores del habla para que se puedan producir diferentes consonantes y vocales. La *Figura 3* muestra la arquitectura implementada.

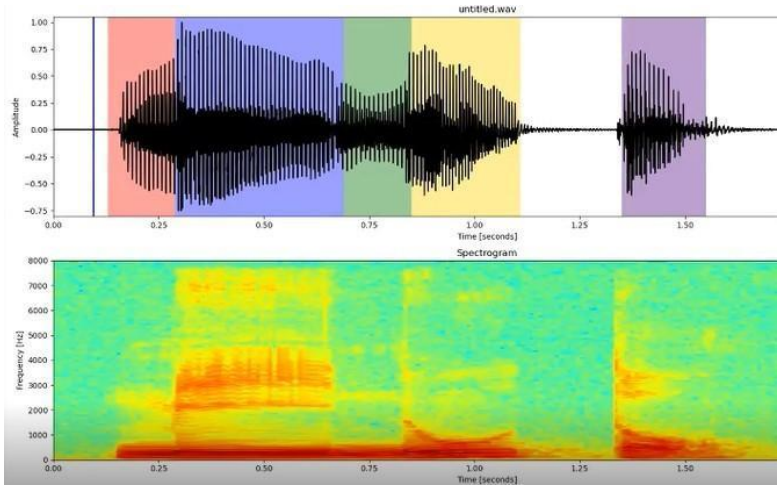

***Figura 2**- Extracción de fonemas en la palabra minuto*

6.3.1.3 MÁQUINA DE SOPORTE VECTORIAL

Podemos definir un método SVM como un conjunto de algoritmos de aprendizaje supervisado desarrollados por Vladimir Vapnik y su equipo, y son utilizados generalmente para resolver problemas de clasificación y regresión. Donde la idea general de este algoritmo es aplicar una transformación no lineal φ (x) del espacio de entrada y mapearlo en una característica dimensional superior espacio en el que se utiliza un límite de decisión lineal (centrado en los datos de entrenamiento) para dividir las dos clases. Tal transformación se realiza para simplificar la construcción de la decisión.

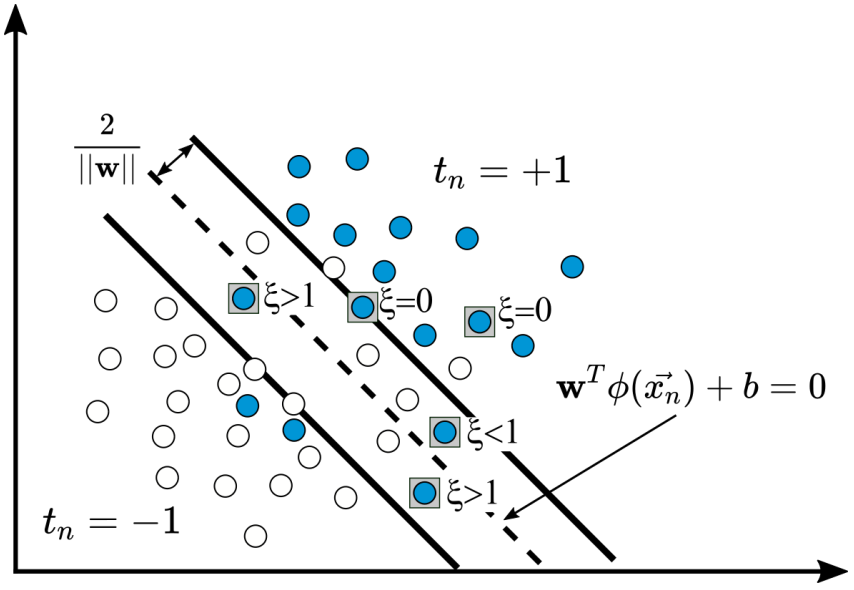

Figura 3 *– Ejemplo de una SVM de margen blando.*

Donde en la *Figura 3* podemos identificar un espacio de características de dos dimensiones donde se puede identificar el hiperplano de separación (línea punteada), dada por la ecuación

$$W^T\phi(\overline{X_n}) + b = 0 \quad \textit{Ecuación 2}$$

Donde de esta ecuación:

- W son los parámetros a ajustar.
- ϕ, es el espacio de características al que se le ha aplicado una transformación **no** lineal
- b es el intercepto de las abscisas

Del mismo modo se tiene la clase $t_n = +1$ como referencia, por ello las variables ξ toma valores de cero si encuentran en el margen correcto, para los puntos dentro de este margen adquirira valores entre $0 < \xi \leq 1$, mientras que si se encuentra en el margen equivocado $\xi > 1$, y de esta manera se busca maximizar el margen mientras se penalizan los puntos de datos para los cuales $\xi > 1$.

6.3.2 APRENDIZAJE DE MÁQUINA

6.3.2.1 GRÁFICA DE RADAR

Un gráfico de radar es una herramienta visual informativa en la que se comparan múltiples variables (tres o más) en un plano bidimensional. Para ello, crearemos diferentes ejes que salen de un punto central común. En la mayoría de los casos, todos los ejes se distribuyen de forma equitativa y se dibujan uniformemente entre sí. A veces, los ejes también están conectados entre sí para formar diferentes cuadrículas que nos facilitan el trazado del gráfico de araña. Los gráficos de radar son considerados como una mejor alternativa a los gráficos de columnas, ya que pueden representar múltiples variables sin problemas. Lo ideal es que un gráfico de araña (radar) se pueda usar en cualquier situación en la que se necesite representar información multivariable en un plano 2D.

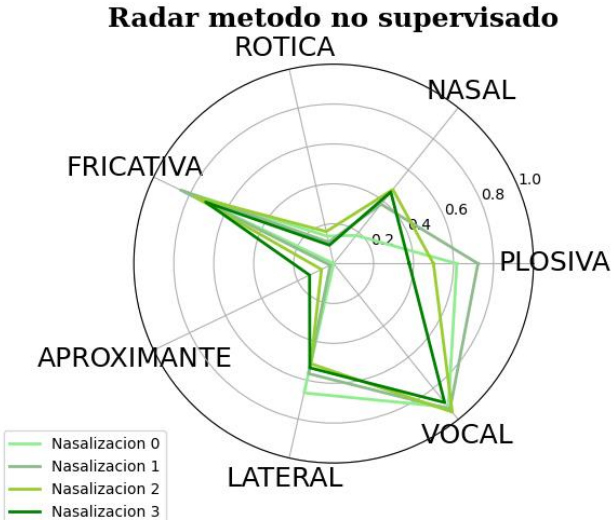

Figura 4*- Ejemplo de una gráfica de radar.*

6.3.2.2. REDES NEURONALES

Las redes neuronales artificiales son un modelo inspirado en el funcionamiento del cerebro humano. Está formado por un conjunto de nodos conocidos como neuronas artificiales que están conectadas y transmiten señales entre sí. Estas señales se transmiten desde la entrada hasta generar una salida.

El objetivo principal de este modelo es aprender modificándose automáticamente a sí mismo, de forma que puede llegar a realizar tareas complejas que no podrían ser realizadas mediante la clásica programación basada en reglas.

Como se ha mencionado, el funcionamiento de las redes se asemeja al del cerebro humano. Las redes reciben una serie de valores de entrada y cada una de estas entradas llega a un nodo llamado neurona. Las neuronas de la red están a su vez agrupadas en capas que forman la red neuronal. Cada una de las neuronas de la red posee a su vez un peso, un valor numérico, con el que modifica la entrada recibida. Los nuevos valores obtenidos salen de las neuronas y continúan su camino por la red.

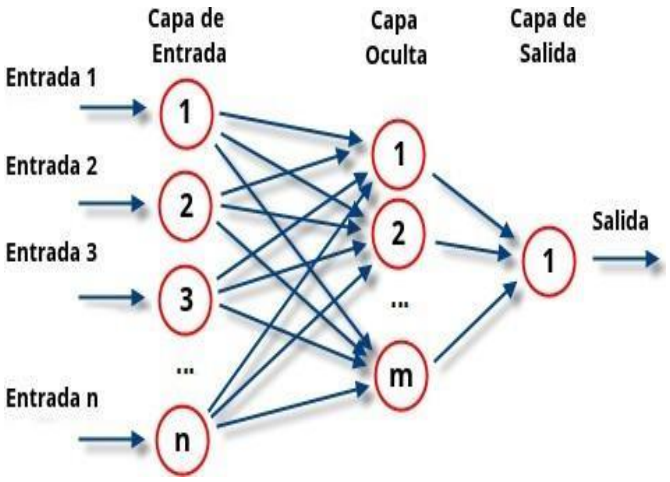

***Figura 5**. Ejemplo del diagrama de una red neuronal*

6.3.2.2.1. REDES NEURONALES RECURRENTES

Las redes neuronales recurrentes (RNN) son una clase de aprendizaje profundo basada en los trabajos de David Rumelhart en 1986. Las RNN son conocidas por su capacidad para procesar y obtener información de datos secuenciales. Por lo tanto, el análisis de video, la subtitulación de imágenes, el procesamiento del lenguaje natural (PLN) y el análisis de la música dependen de las capacidades de las redes neuronales recurrentes. A diferencia de las redes neuronales artificiales ya vistas, que asumen la independencia entre los datos de entrada, las RNN capturan activamente sus dependencias secuenciales y temporales.

Uno de los atributos más definitorios de las RNN es la compartición de sus parámetros. Sin compartir parámetros, el modelo asignaría parámetros únicos para representar a cada dato en una secuencia y, por lo tanto, no podría realizar inferencias sobre secuencias de longitud variable.

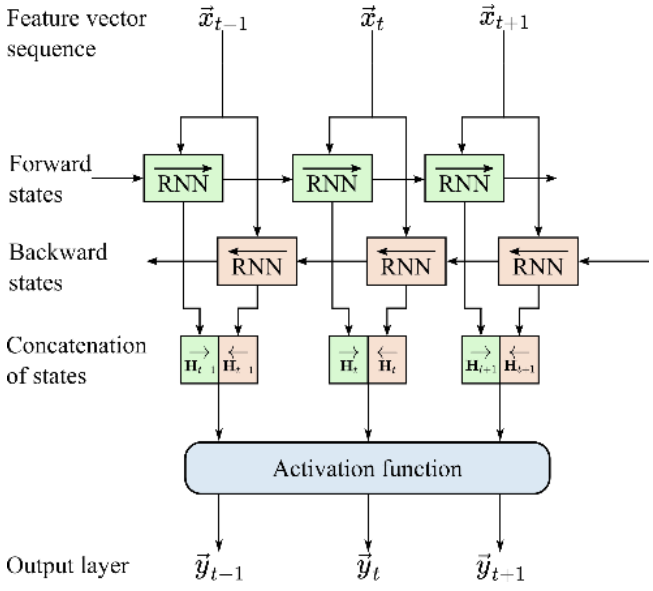

***Figura 6**- Ejemplo de red neuronal recurrente.*

6.3.2.2.2 ALINEAMIENTO FORZADO

El alineamiento forzado se refiere al proceso de sincronizar una señal de voz con su respectiva transcripción ortográfica. Mediante este proceso, es posible identificar los instantes de tiempo en los que ciertas palabras (o fonemas) fueron pronunciados. Para este proceso se requiere de un sistema de reconocimiento de voz utiliza un motor de búsqueda junto con un modelo acústico y de lenguaje que contiene un conjunto de posibles palabras, fonemas o algún otro conjunto de datos para hacer coincidir los datos del habla a la expresión hablada correcta. El motor de búsqueda procesa las características extraídas de los datos del habla para identificar las ocurrencias de las palabras, fonemas o cualquier conjunto de datos que esté equipado para buscar y devuelve los resultados, el alineamiento forzado, es similar a este proceso, pero difiere en un aspecto importante. En lugar de recibir un conjunto de posibles palabras para buscar, el motor de búsqueda recibe un extraído de lo que se habla en los datos de voz. Luego, el sistema alinea los datos transcritos con los datos de voz, identificando qué segmentos de tiempo en los datos de voz corresponden a palabras particulares en los datos de transcripción.

7. METODOLOGÍA

Para la investigación que se plantea se definió una metodología de evaluación y caracterización, la cual se aplica a todos los datos (Grabaciones de niños con CLP) que hacen parte del banco de datos, se va a analizar la precisión con la que los pacientes producen ciertos fonemas.

Para eso se hará uso de una red neuronal recurrente, que ha sido pre-entrenada con una base de datos con grabaciones en español e inglés. La red neuronal transformará una señal de voz en una secuencia de probabilidades de ocurrencia de fonemas en una grabación de voz.

Esta metodología se diseñó de tal manera que permite evaluar la pronunciación de diferentes fonemas, palabras, frases que sean potencialmente, útiles en la identificación de nasalización en niños, por medio de diferentes algoritmos de aprendizaje profundo que sean supervisados y semi-supervisado, como lo son las redes neuronales recurrentes, donde dichas redes neuronales serán implementadas utilizando la librería Pytorch (https://pytorch.org/) y webmaus para el alineamiento forzado.

Se consideraron frases en dos idiomas el español y el inglés donde en español se tienen 88 grabaciones de voz de niños con CLP, cada uno de niños leyó un máximo de seis frases mientras en inglés se tiene 44 grabaciones de niños con CLP cada uno de los niños leyó un total de tres frases, donde todas las frases del proyecto son las siguientes:

Base de datos en español

- Carlos coge su pelota (Frase 1)
- El gato toma leche (Frase 2)
- La llave de la casa (Frase 3)
- La silla es café (Frase 4)
- Susi come sopa (Frase 5)
- Tomás toca tambor (Frase 6)

Base de datos en inglés

- Big red truck (Frase 7)
- Chase the chickens (Frase 8)
- Chocolate chip cookie (Frase 9)

Adicionalmente el nivel de nasalización de los niños fue evaluado por un experto, don- de la distribución de grupos de niños de acuerdo al nivel de hipernasalidad es de las grabaciones en español:

- 20 niños con hipernasalidad tipo 0
- 28 niños con hipernasalidad tipo 1
- 25 niños con hipernasalidad tipo 2
- niños con hipernasalidad tipo 3

Mientras que de las grabaciones en inglés se tiene:

- 15 niños para el control
- 28 niños con hipernasalidad

La primera actividad a realizar es teniendo estas grabaciones, primero se realiza un análisis de probabilidad de ocurrencia de un tipo de fonema por persona y por cada uno de los grupos de nasalización que se tienen, posterior a tener estos resultados la idea es agrupar los fonemas específicos teniendo en cuenta la transliteración (Contexto en que son dichos estos fonemas), para se realiza alineamiento forzado haciendo uso de la herramienta WebMAUS, con esto lo que se pretende es separar todas las vocales de cada una de las oraciones y con ella todos los posteriors encontrados con anterioridad con ayuda del Forced Alignment, y del mismo modo se realizará en análisis con las fricativas, nasales, plosivas, laterales de esta manera se realizará un análisis por grupo fonémico partiendo de la hipótesis que siempre se tendrá la transliteración, con estos resultados se creará el mapa fonético.

Donde para esto es necesario conocer lo que es la fonética a la luz de la hipernasalidad, debido a que este fenómeno afecta a todos los fonemas (nasales, laterales fricativas plosivos entre otros) y para realizar un análisis efectivo es necesario conocer de qué manera y en que escenarios se ven

afectados cada uno de estos. Un segundo análisis, es después de tener la agrupación de los fonemas, es evaluar qué pasa según el contexto fonético, de esta manera se planea obtener resultados la detección de hipernasalidad en pacientes con CLP, con diferentes tipos de clasificadores.

8. EXPERIMENTOS Y RESULTADOS

Los experimentos realizados a continuación fueron llevados a cabo con ayuda de una red neuronal recurrente, que ha sido pre-entrenada, donde se harán uso de los idiomas inglés y español, esto con el fin de probar que hay ciertos sonidos de la voz que se ven afectados por la nasalización y es posible cuantificar la precisión con la que se producen esos sonidos.

8.1.BASE DE DATOS EN ESPAÑOL

8.1.1. IDENTIFICACIÓN DE NASALIZACIÓN

La primera actividad realizada fue aplicar la red neuronal mediante:

- Un método no supervisado: es la obtención de los posterios de todos los fonemas independiente de que este fonema se encuentre en la oración, es decir nos muestra como lo hizo para todas las clases, aunque en esa oración no haya todos los fonemas.

- Un método semi-supervidado: en esta ocasión es la obtención de los posterios sólo tomando los fonemas que se encuentran dentro de la oración.

- Método supervisado: para este se hizo uso del alineamiento forzado y se obtuvo los posteriros en el momento en el que se pronuncia el fonema.

Donde los fonemas analizados con ayuda de esta red, los podemos observar en la *Figura 7*, todos estos fonemas se encuentran a lo largo de todas las frases analizadas lo que nos permite tener una idea del comportamiento de cada uno de ellos.

c	Dimension	Class	Phonemes
0	-	Silence	-
1	Manner	Stop	/p/, /t/, /k/, /b/, /d/, /g/
2		Nasal	/n/, /m/, /ŋ/
3		Trill	/r/, /ɾ/
4		Fricative	/s/, /ʃ/, /z/, /f/
5		Approximants	/j/
6		Lateral	/l/
7		Vowel	/a/, /e/, /i/, /o/, /u/

Figura 7. *Fonemas analizados.*

En la *Figura 8*, se muestra la máxima probabilidad de un fonema en cada uno de los grupos de nasalización, en esta figura contamos con tres columnas donde cada columna es método supervisado, semi-supervisado, y no supervisado respectivamente, y las Filas son las frases de la 1 la 6 respectivamente.

Analizando la *Figura 8* se puede evidenciar que los fonemas a posteriori son relevantes para el análisis de la nasalización por:

- En las frases Carlos coge su pelota primera Fila, la silla es café (Fila 4), la llave de la casa (Fila 2), estas oraciones no poseen un fonema nasal, pero con el método no supervisado se puede evidenciar que el sistema si detecta sonidos nasales esto puede ser debido a que el sistema se confunde porque detecte algunos sonidos que pueden ser nasales y mucho más evidentes en las clases dos y tres, y esto es de bastante utilidad debido a que nos indica que el sistema es bueno a la hora de detectar nasalización.

- Del mismo modo se puede evidenciar en todas las oraciones que los métodos semi supervisado y supervisado son bastantes similares, aunque se puede evidenciar algunas diferencias bastante notables, por ejemplo, en la Fila 1 en el método semi-supervisado a una mayor probabilidad para los posterios de las plosivas que si se encuentra dentro de la oración en especial para la clase uno de nasalización, del mismo modo en la Fila 3 se ve una mayor probabilidad para los postrios de las lateral en el método semi- supervisado que, en el método supervisado, esto nos da una idea de que dentro del sistema se encuentran

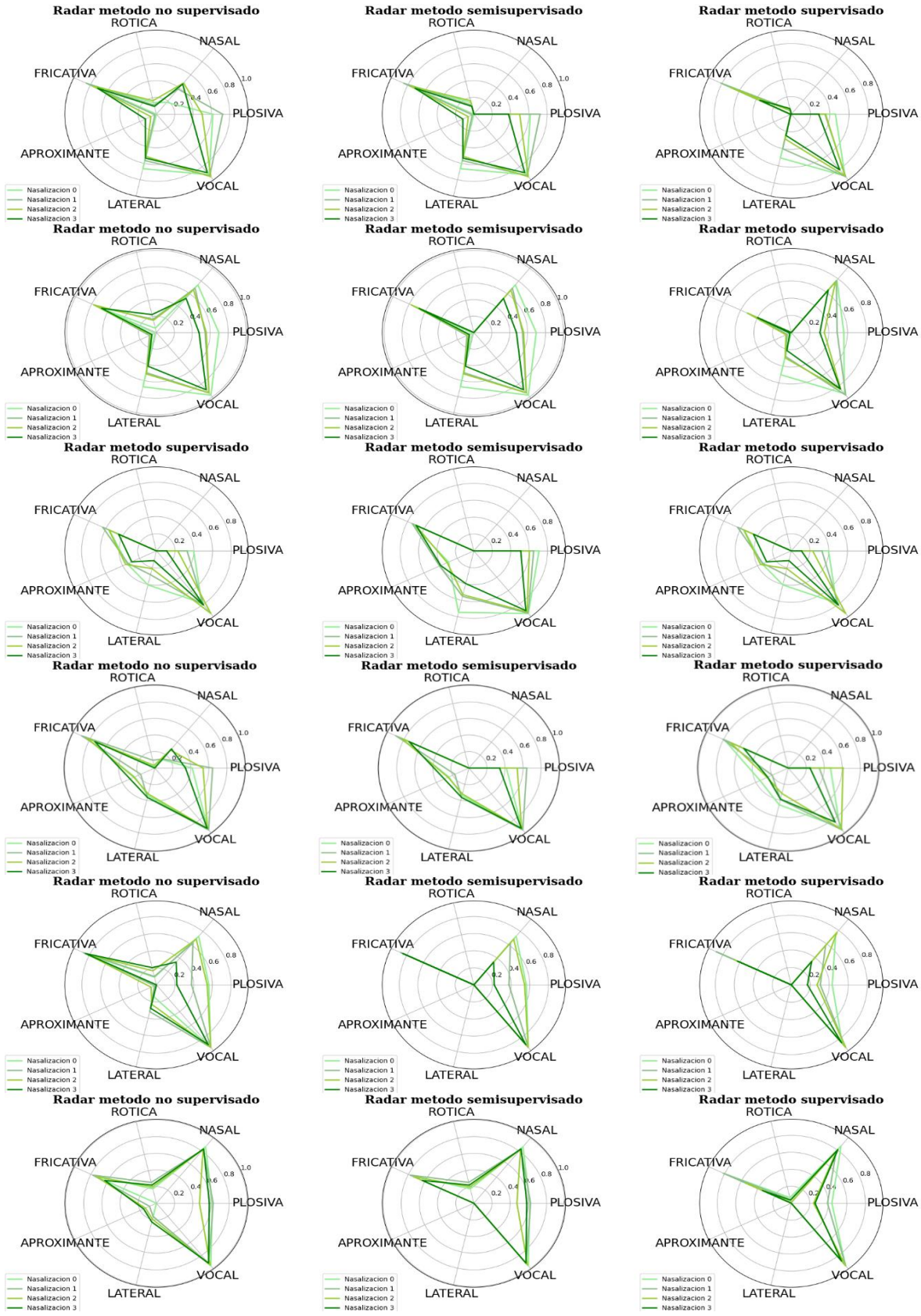

Figura 8 *– Comparación resultados de base de datos en español*

mayores diferencias en los que son las plosivas, laterales, pero aun así el sistema es bastante confiable.

- Del sistema también se puede evidenciar de las oraciones que poseen roticas (r), como lo son la Fila 1 y Fila 6, este fonema es difícil de evaluar debido es difícil de pronunciar para los niños, aunque también el sistema tiene una muy baja tasa de acierto para este fonema, es por ello que a pesar de que este se encuentra en la oración, la probabilidad mostrada en los radares es bastante baja, lo mismo sucede con los laterales.

Para tener una idea mucho más clara de estos resultados que se muestran en la columna tres *Figura 8* (método supervisado) a continuación, se exponen las probabilidades de los fonemas con su respectiva desviación estándar, estas probabilidades se obtuvieron por medio del alineamiento forzado donde se toman las grabaciones en los diferentes grupos de nasalización (nasalización 0, nasalización 1, nasalización 2, y nasalización 3), por cada grabación se analizan los fonemas presentes en el momento de tiempo en que son dichos, se toma la probabilidad por oración y posteriormente se calcula la probabilidad por grupo, donde con estos datos se puede ver de una manera más analítica la información graficada, es decir que tan lejos o que tan cercanos está un grupo del otro por medio del método supervisado.

- Calor coge su pelota.

Tabla 1. *Tabla de probabilidades Carlos coje su pelota*

	Aproximante	**Fricativa**	**Lateral**	**Plosiva**	**Rotica**	**Vocal**
Nasalización 0	0.006±0.014	0,799±0.304	0.507±0.316	0.458±0.201	0.055±0.186	0.894±0.121
Nasalización 1	0.010±0.03	0.812±0.286	0.401±0.286	0.349±0.252	0.052±0.138	0.906±0.0901
Nasalización 2	0.004±0.006	0.705±0.314	0.288±0.198	0.356±0.178	0.039±0.142	0.892±0.100
Nasalización 3	0.005±0.014	0.363±0.315	0.247±0.238	0.284±0.272	0.065±0.155	0.807±0.144

- El gato toma leche

Tabla 2. *Tabla de probabilidades El gato toma leche*

	Aproximante	**Fricativa**	**Lateral**	**nasal**	**plosiva**	**Vocales**
Nasalización 0	0.000±0.001	0.528±0.440	0.503±0.313	0.780±0.344	0.560±0.368	0.910±0.086
Nasalización 1	0.047±0.187	0.453±0.435	0.284±0.288	0.770±0.368	0.489±0.351	0.931±0.08
Nasalización 2	0.061±0.142	0.516±0.426	0.301±0.228	0.757±0.371	0.352±0.269	0.848±0.180
Nasalización 3	0.017±0.058	0.399±0.435	0.214±0.274	0.632±0.384	0.305±0.309	0.839±0.183

- La llave de la casa

Tabla 3. *Tabla de probabilidades La llave de la casa*

	Aproximante	**Fricativa**	**Lateral**	**Plosivas**	**Vocal**
Nasalización 0	0.346±0.372	0.558±0.405	0.406±0.265	0.397±0.279	0.776±0.229
Nasalización 1	0.3367±0.356	0.6379±0.438	0.284±0.281	0.327±0.342	0.808±0.201
Nasalización 2	0.3648±0.3963	0.564±0.423	0.2121±0.220	0.228±0.157	0.933±0.057
Nasalización 3	0.2935±0.3893	0.448±0.424	0.1169±0.178	0.108±0.148	0.8065±0.115

- La silla es café

***Tabla 4**. Tabla de probabilidades la silla es café*

	Aproximante	**Fricativa**	**Lateral**	**stop**	**Vocal**
Nasalización 0	0.348±0.380	0.786±0.177	0.455±0.410	0.469±0.375	0.9217±0.077
Nasalización 1	0.198±0.324	0.807±0.165	0.391±0.413	0.345±0.415	0.96140±0.065
Nasalización 2	0.260±0.375	0.716±0.210	0.316±0.375	0.609±0.400	0.9485±0.065
Nasalización 3	0.253±0.405	0.550±0.220	0.387±0.458	0.242±0.332	0.8387±0.120

- Susi come sopa

***Tabla 5**. Tabla de probabilidades susi come sopa*

	Fricativa	**Nasal**	**Plosivas**	**vocales**
Nasalización 0	0.840±0.187	0.751± 0.395	0.431± 0.337	0.861± 0.158
Nasalización 1	0.877± 0.146	0.580± 0.451	0.305± 0.292	0.914± 0.091
Nasalización 2	0.661± 0.235	0.773± 0.378	0.271±0.267	0.919± 0.089
Nasalización 3	0.6279± 0.251	0.344± 0.382	0.170±0.267	0.848± 0.123

- Tomas toca tambor

***Tabla 6**. Tabla de probabilidades Tomas toca tambor*

	Fricativa	**Nasal**	**plosivas**	**rotica**	**vocales**
Nasalización 0	0.809±0.331	0.861±0.185	0.438±0.256	0.067±0.216	0.939±0.073
Nasalización 1	0.817±0.276	0.777±0.292	0.391±0.219	0.043±0.183	0.943±0.071
Nasalización 2	0.548±0.411	0.784±0.216	0.241±0.161	0.009±0.019	0.937±0.065
Nasalización 3	0.346±0.343	0.807±0.230	0.256±0.223	0.036±0.114	0.870±0.179

8.1.2 MÉTODOS DE CLASIFICACIÓN

Para la clasificación se decide trabajar con la máquina de soporte vectorial (SVM) para multiclase, teniendo en cuenta que en trabajos anteriores se utilizó este mismo método de clasificación, con MFCC contentado con GFCC aplicándole un VAD los resultados obtenidos de ese experimento fueron los que se pueden ver en la *tabla 7*

***Tabla 7**. SVM aplicando MFCC y GFCC*

		SVM				
	Tarea	**C**	**Gamma**	**Precisión**	**Recall**	**F1 score**
MFCC+GFCC Con VAD	Frase 1	10.0	0.01	0.357	0.318	0.314
	Frase 2	10.0	0.1	0.77	0.25	0.118
	Frase 3	10.0	0.1	1.079	1.25	0.121
	Frase 4	10.1	0.01	0.245	0.231	0.226
	Frase 5	1000	0.0001	0.336	0.34	0.335
	Frase 6	10.0	0.01	0.344	0.315	0.30

Es por ello que, aplicando la misma máquina de soporte vectorial a los datos obtenidos del método supervisado, es decir de las probabilidades de los fonemas posteríos, al mismo tiempo que aplicándola a la LLRPh (razón logarítmica de verosimilitud del promedio posterior) esto debido a que la probabilidad son números entre 0 y 1 esto nos puede causar algunos problemas al aplicar este método, y también se decidió aplicarlo a la concatenación de ambos y los resultados obtenidos son los que se pueden evidenciar en la *tabla 8.*

De la *tabla 8* de se puede evidenciar que los datos obtenidos con LLRPh, son ligeramente mejores que los obtenidos con los fonemas, y del mismo modo la concatenación de estos es un poco mejor que los anteriores y estos tres resultados son mejores que los obtenidos solo con MFCC concatenado con GFCC, lo que nos da a entender que el sistema es mejor si se trabaja por medios supervisados evaluando los fonemas posteriores.

***Tabla 8**. SVM aplicando a los fonemas posteriors*

	Tarea	SVM		Precisión	Recall	F1 score
		C	Gamma			
Fonemas	Frase 1	1.0	0.01	0.079	0.241	0.119
	Frase 2	**1000.0**	**0.1**	**0.329**	**0.320**	**0.321**
	Frase 3	10.0	0.001	0.029	0.715	0.042
	Frase 4	1000.0	0.001	0.300	0,297	0.297
	Frase 5	**1.0**	**0.1**	**0.496**	**0.478**	**0.480**
	Frase 6	100.0	0.01	0.254	0.257	0.231
LLRPh	Frase 1	10.0	1.0	0.235	0.271	0.245
	Frase 2	**1.0**	**0.1**	**0.381**	**0.300**	**0.324**
	Frase 3	10	0.001	0.073	0.223	0.111
	Frase 4	1000.0	0.1	0.302	0.285	0.290
	Frase 5	**10.0**	**0.1**	**0.438**	**0.422**	**0.427**
	Frase 6	**100**	**0.1**	**0.346**	**0.316**	**0.327**
Fonemas + LLRPh	**Frase 1**	**10000.0**	**0.001**	**0.429**	**0.423**	**0.421**
	Frase 2	100.0	0.01	0.268	0.271	0.269
	Frase 3	1.0	0.01	0.023	0.035	0.032
	Frase 4	**1000.0**	**0.0001**	**0.276**	**0.257**	**0.265**
	Frase 5	**10.0**	**0.01**	**0.469**	**0.455**	**0.455**
	Frase 6	**100.0**	**0.01**	**0.281**	**0.287**	**0.282**

Teniendo como base que la concatenación de los MFCC y los GFCC nos arrojan los mejores resultados se decide realizar el experimento de la máquina sobre la concatenación de la matriz obtenida de los MFCC con los GFFF y las combinaciones vistas en la *tabla 8.*

A pesar de que se podría pensar que al concatenar MFCC, GFCC y probabilidades de fonemas que se tiene, se obtendrían unos muy buenos resultados peros como podemos observar en la *tabla 9* no mejoran los resultados, por el contrario se puede ver que los resultados desmejoran bastante, esto se puede deber la cantidad de características crece mucho más que la cantidad de muestras disponibles (grabaciones de voz), lo cual ocasiona que para el algoritmo sea "más difícil" encontrar los patrones que separan una clase de la otra, debido a que en el momento en que se tienen demasiadas características, las

observaciones se vuelven más difíciles de agrupar; esto es debido a que demasiadas dimensiones hacen que cada observación en su conjunto de datos parezca equidistante de todas las demás (Curse of dimensionality).

Tabla 9. *SVM concatenado con MFCC y GFCC*

		SVM				
	Tarea	**C**	**Gamma**	**Precisión**	**Recall**	**F1 score**
MFCC+CFCC +Fonemas	**Frase 1**	**100.0**	**0.001**	**0.368**	**0.365**	**0.363**
	Frase 2	10.0	0.1	0.077	0.250	0.118
	Frase 3	0.01	0.01	0.079	0.250	0.121
	Frase 4	**10.0**	**0.1**	**0.326**	**0.301**	**0.310**
	Frase 5	**10.0**	**0.1**	**0.402**	**0.390**	**0.393**
	Frase 6	1.0	0.01	0.136	0.121	0.180
MFCC+CFCC +LLRPh	Frase 1	0.01	0.0001	0.092	0.250	0.135
	Frase 2	10.0	0.01	0.272	0.2405	0.240
	Frase 3	10.0	0.1	0.079	0.250	0.121
	Frase 4	10.0	0.1	0.805	0.250	0.122
	Frase 5	**10.0**	**0.001**	**0.379**	**0.365**	**0.361**
	Frase 6	1.0	0.01	0.124	0.177	0.134
MFCC+GFCC+Fonemas + LLRPh	Frase 1	0.01	0.0001	0.092	0.25	0.135
	Frase 2	10.0	0.001	0.290	0.259	0.263
	Frase 3	10.0	0.1	0.079	0.25	0.121
	Frase 4	1.0	0.01	0.124	0.209	0.156
	Frase 5	**10.0**	**0.01**	**0.436**	**0.419**	**0.422**
	Frase 6	1.0	0.01	0.144	0.195	0.154

Teniendo los datos anteriores donde se puede evidenciar un aumento en los resultados de clasificación se decide realizar un análisis de PCA al mejor de los resultados obtenidos la oración "***Susi come sopa***" donde a nivel gráfico tenemos:

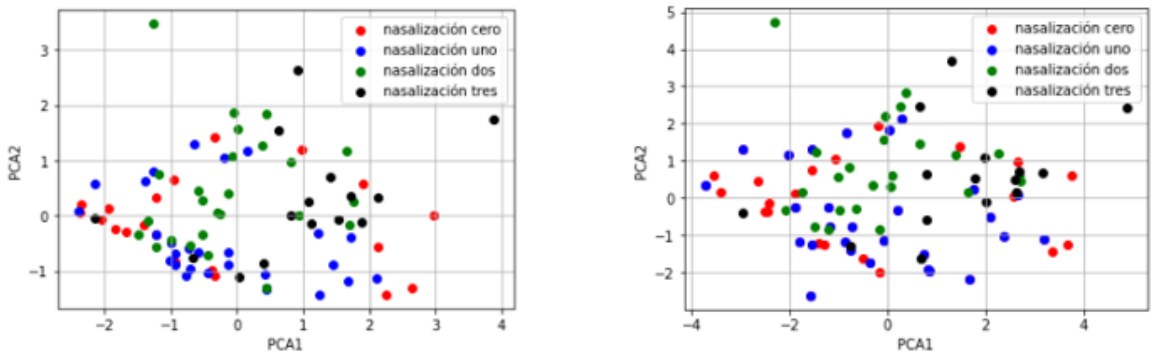

Figura 9*- Análisis de PCA para la oración Susi come sopa*

En la *Figura 9* tenemos todos los tipos de nasalización, en la parte izquierda se tienen graficados los datos de los fonemas, mientras que en la parte derecha se tienen graficados los datos de la LLRPh, de este análisis se puede evidencia que no se alcanza a ver mucha diferenciación, entre los diferentes grupos, es por ello que este análisis no es de gran relevancia a lo largo de este trabajo.

8.2.BASE DE DATOS INGLÉS HINDÚ

Para este caso se cuenta con una base de datos que fue compartida por colegas del International Institute of Information Technology in Hyderabad, India. Específicamente Prof. Anil Kumar Vuppala. Donde de esta base de datos se contaba con una gran cantidad de grabaciones, pero teniendo en cuenta la finalidad de este trabajo es decir se pretende realizar un análisis por medio de alineamiento forzado, se decidió trabajar con las oraciones que más personas dijeron, es decir las oraciones que tenían mayor cantidad de grabaciones las cuales fueron las oraciones de la 7 a la 9.

8.2.1. IDENTIFICACIÓN DE NASALIZACIÓN

Como en los experimentos de la base de datos en español lo primeros que se realizó fue aplicar la red neuronal por medio de los tres métodos antes mencionados (no supervisado, semi-supervisado, supervisado).

Donde la disposición de las columnas son los métodos en el orden antes mencionado mientras que las Filas se encuentra desde la oración 7 a la oración 9, esta base de datos en inglés cuenta con grabaciones de control y grabaciones de hipernasalidad, es decir para este caso solo se cuenta con dos grupos de estudio.

En la *Figura 10,* en la Fila 1 se puede evidenciar que esta oración con el método no supervisado, nos arroja probabilidades en todos los grupos fonéticos a pesar de que esta oración solo cuenta con tres de ellos, esto se debe a que la red neuronal está entrenada con inglés americano (EE. UU.), y del mismo modo el alineamiento forzado se realizó con este inglés, mientras que las grabaciones del estudio son grabaciones en inglés hindú, es por ello que se puede ver que el sistema presenta fallos debido a que a pesar de que ambos son inglés, se presentan diferencias en la pronunciación, que es la parte fundamental a la hora de realizar este tipo de análisis, y esto se puede ver en las otras dos oraciones.

Pero a pesar de esta falla para este tipo de grabaciones también se puede evidenciar que los resultados obtenidos por medio de los métodos semi-supervisado y supervisado son muy similares del mismo modo como sucedió con las grabaciones en español.

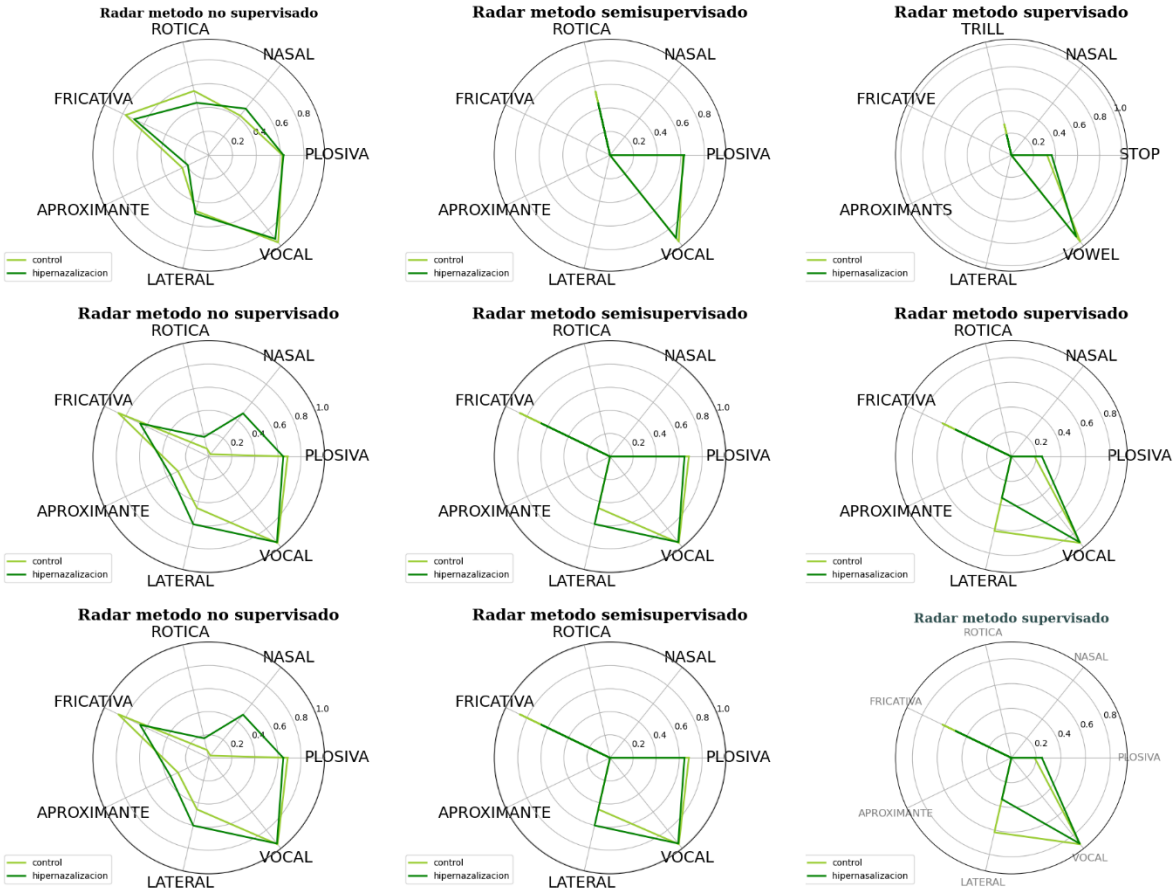

Figura 10 *Comparación resultados base de datos en inglés*

Para tener una idea mucho más clara de estos resultados que se muestran en la columna tres *Figura 10* (método supervisado) a continuación, se exponen las probabilidades de los fonemas con su respectiva desviación estándar, estas probabilidades se obtuvieron por medio del alineamiento forzado donde se toman las grabaciones en los diferentes grupos (es decir grabaciones de control, y grabaciones con hipernasalidad), por cada grabación se analizan los fonemas presentes en el momento de tiempo en que son dichos, se toma la probabilidad por oración y posteriormente se calcula la probabilidad por grupo, donde con

estos datos se puede ver de una manera más analítica la información graficada, es decir que tan lejos o que tan cercanos está un grupo del otro por medio del método supervisado.

- Big red truck

***Tabla 10**. Tabla de probabilidades Big red truck*

	Plosiva	**Rotica**	**Vocal**
Control	0.323±0.183	0.285±0.231	0.999±0.008
Hiper	0.364±0.2527	0.187±0.255	0.943±0.141

- Chase the chickens

***Tabla 11**. Tabla de probabilidades chase the chickens*

	Fricativa	**Nasal**	**plosiva**	**Vocal**
Control	0.823±0.159	0.017±0.055	0.034±0.070	0.837±0.070
Hiper	0.653±0.271	0.172±0.339	0.164±0.174	0.890±0.121

- Chocolate chip cookie

***Tabla 12**. Tabla de probabilidades Chocolate chip cookie*

	Fricative	**Lateral**	**Plosiva**	**Vocal**
Control	0.617±0.251	0.188±0.314	0.188±0.151	0.893±0.111
Hiper	0.499±0.368	0.342±0.419	0.247±0.209	0.885±0.132

8.2.2 MÉTODOS DE CLASIFICACIÓN

En trabajos anteriores no se tenía un análisis de las oraciones en inglés, el primer análisis de clasificación que se realizó con estas oraciones son las MFCC y las GFCC con VAD, donde los resultados obtenidos de este primer análisis se ven reflejado en la tabla 13, aunque se puede ver que estos resultados son bastante buenos debido a que estos (MFCC y GFCC) son "independientes" del lenguaje (Es decir no son sensibles al tipo de inglés sea inglés Hindú o inglés estadounidenses) al fin y al cabo estos datos poseen información comprimida de la distribución de energía espectral.

Tabla 13. *SVM aplicado con MFCC y GFCC*

	Tarea	SVM		Precisión	Recall	F1 score
		C	Gamma			
MFCC+GFCC	Frase 7	100.0	0.001	0.902	0.902	0.902
	Frase 8	1.0	0.001	0.845	0.843	0.840
	Frase 9	1.0	0.01	0.907	0.906	0.909

Posterior a este primer análisis, se procede a realizar la clasificación a los datos obtenidos por medio del alineamiento forzado, es estos se puede ver que los resultados desmejoran considerablemente, pero esto es lo esperado debido a que a pesar de que el sistema no está entrenado para estos datos, se ve que puede distinguir algunos de los fonemas que se tiene una pronunciación en ambos ingleses.

Tabla 14. *SVM aplicado a los fonemas posteriors*

	Tarea	SVM		Precisión	Recall	F1 score
		C	Gamma			
Fonemas	Frase 7	100.0	1.0	0.624	0.612	0.614
	Frase 8	10.0	0.1	0.642	0.634	0.637
	Frase 9	100.0	1.0	0.624	0.612	0.614
LLRPh	Frase 7	1.0	1000.0	0.310	0.500	0.038
	Frase 8	1.0	0.055	0.650	0.598	0.590
	Frase 9	1.0	1000.0	0.318	0.500	0.389
Fonemas + LLRPh	Frase 7	1.0	100.0	0.833	0.522	0.511
	Frase 8	1.0	0.01	0.309	0.500	0.382
	Frase 9	1.0	100.0	0.833	0.562	0.511

Y finalmente se realiza en método de clasificación con la concatenación de las dos pruebas antes realizadas, en este caso podemos ver que el sistema en muchas de las oraciones nos da resultados demasiado buenos y en otros no es capaz de realizar ningún tipo de reconocimiento, , esto se puede deber al mismo motivo presentado para la base de datos de español el cual es Curse of dimensionality.

Tabla 15. *SVM concatenado con MFCC y GFCC*

	Tarea	SVM		Precisión	Recall	F1 score
		C	Gamma			
MFCC+CFCC +Fonemas	Frase 7	1000.0	0.0001	0.902	0.902	0.902
	Frase 8	1.0	0.1	0.911	0.887	0.896
	Frase 9	1.0	1000.0	0.310	0.500	0.038
MFCC+CFCC + LLRPh	Frase 7	100.0	0.001	0.902	0.902	0.902
	Frase 8	1.0	0.01	0.911	0.887	0.896
	Frase 9	1.0	1000.0	0.310	0.500	0.038
MFCC+GFCC+fonemas + LLRPh	Frase 7	100.0	0.001	0.902	0.902	0.902
	Frase 8	1.0	0.001	0.3095	0.500	0.382
	Frase 9	1.0	1000.0	0.310	0.500	0.038

Debido a que estos datos no nos presentan información relevante para nuestra investigación por los motivos mencionados anteriormente, a esta base de datos no se le realiza un análisis con PCA.

8.3. ANÁLISIS FONÉMICO

Al momento de realizar un análisis fonémico se decide trabajar con la oración que mejor resultado arrojó a lo largo de esta investigación y esta fue "***Susi come sopa***", de esta oración se analizó la transición (Consonante-vocal) de los fonemas que un mejor desempeño presentaron los cuales fueron las fricativas y las vocales donde los resultados obtenidos son los siguientes:

Tabla 16. *Análisis fonémico de las fricativas en la oración susi come sopa*

	Fricativa			
	Consonante	**Consonante_nasal**	**Vocal**	**Vocal_nasal**
Nasalización 0	0.8403	0.046	0.9984	0.12
Nasalización 1	0.8775	0.074	0.9472	0.182
Nasalización 2	0.6619	0.093	0.7886	0.245
Nasalización 3	0.6279	0.098	0.9908	0.31

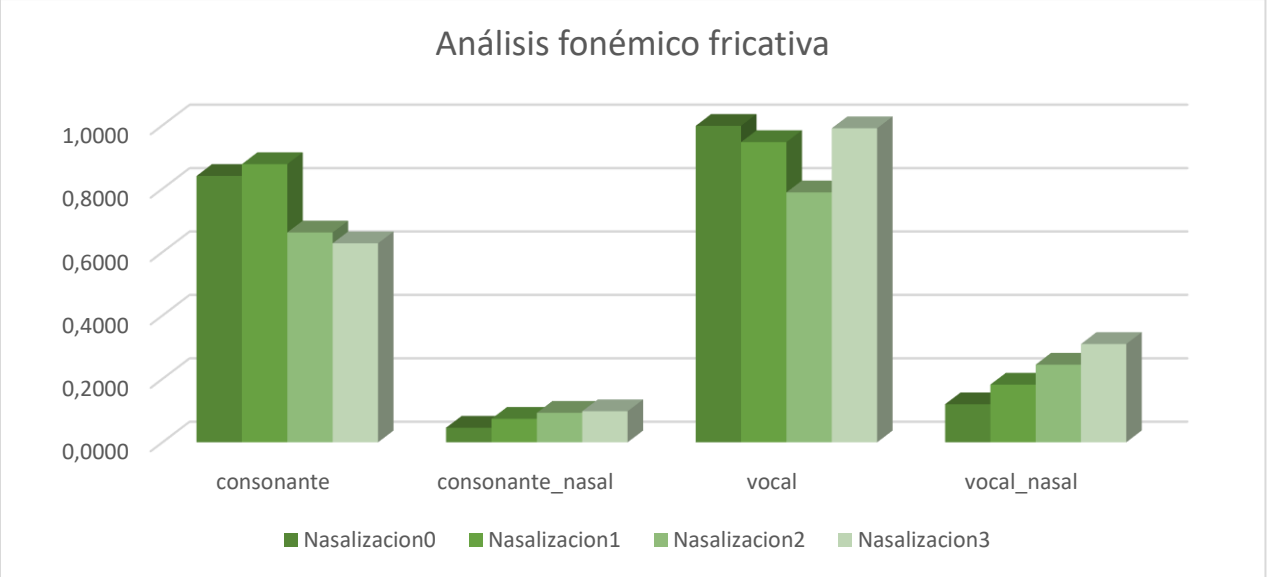

Figura 11 *– Diagrama de Barra análisis fonémico de las fricativas.*

Tabla 17. *Análisis fonémico de las fricativas en la oración susi come sopa*

	Plosivas			
	consonante	**consonante_nasal**	**vocal**	**vocal_nasal**
Nasalización 0	0.4311	0.237	0.967	0.306
Nasalización 1	0.3051	0.275	0.999	0.362
Nasalización 2	0.2715	0.379	0.9972	0.42
Nasalización 3	0.1700	0.485	0.978	0.51

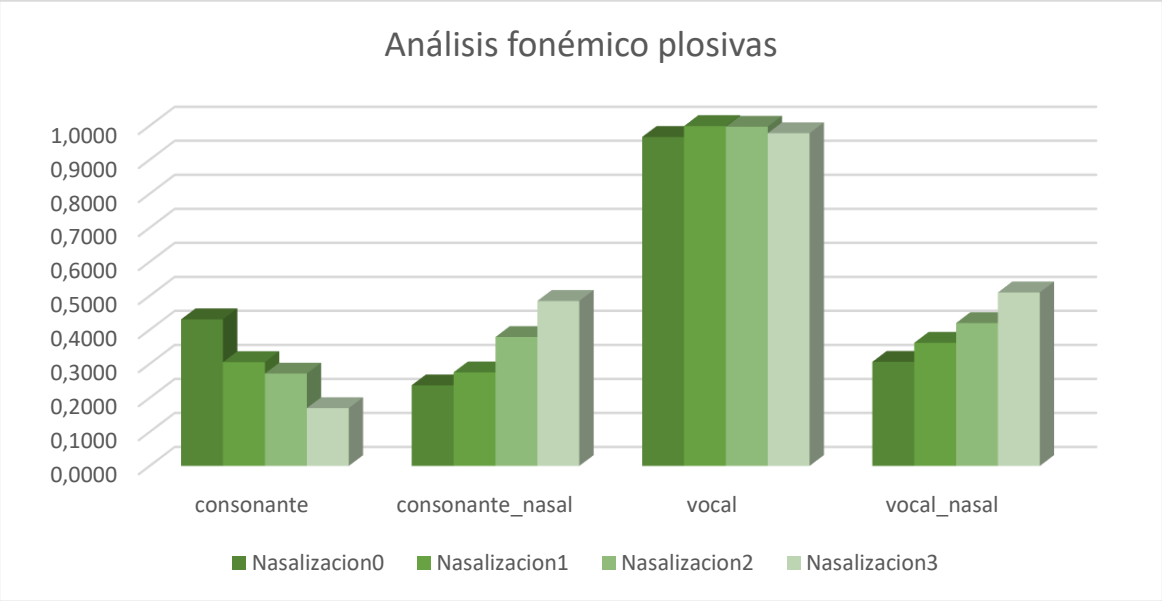

Figura 12 – *Diagrama de Barra análisis fonémico de las plosivas*

De las *Figura 11*, se puede evidenciar que en el fonema se presenta, aunque bajo, un nivel de nasalización es cual es detectado por el sistema, del mismo modo se puede evidenciar en las vocales, y este sonido nasal es un poco más alto que el presente en la consonante esto es debido a que en las vocales hay mayor nasalización porque es más fácil que el sistema confunda una nasal con una vocal. Las nasales se caracterizan por tener una estructura espectral armónica parecida a las vocales. Si hay exceso de nasalización, el sistema se va a confundir más.

A pesar de que el efecto coarticulación es un tema que puede ser de gran utilizad al comento de la cuantificación de los diferentes niveles de nasalización, este tema es algo que va mucho más allá de los alcances de esta investigación, por lo que este análisis es un trabajo que queda a futuro.

9. CONCLUSIONES

- Según los datos obtenidos, cuando no hay un sonido nasal en la frase, aún se puede apreciar probabilidades de un sonido nasal para las clases con mayor nivel de nasalización, las probabilidades de sonidos nasales son mayores que en los otros grupos, lo cual indica que si es posible detectar el nivel de nasalización utilizando un reconocedor automático de fonemas.

- Las metodologías semi-supervidadas y supervisadas son bastante similares entre ellas, eso nos da a entender que el sistema con el que se trabajó, nos da resultados confiables.

- Con la metodología implementada es posible detectar los diferentes niveles de nasalización con un F1 Score de 0.48, para datos con fonemas, mientras que con LLRPh el mejor resultado obtenido es un F1 Score de 0.42, aunque este es un poco menor que el de los fonemas los datos en general de todo el grupo con este logaritmo son mejores que con los fonemas, esto puede deberse a que los datos del grupo de fonemas son probabilidades es decir se encuentran entre 0 y 1 esto puede causar alguna confusión en el sistema.

- Al momento de realizar la clasificación de los resultados obtenidos por medio del alineamiento forzado podemos ver que se mejoraron un poco los resultados obtenidos en pruebas anteriores, es por ello que el análisis de los fonemas posterior es de gran utilidad a la hora de cuantificar los sonidos nasales, debido a que nos permite identificar estos sonidos en cualquiera de los fonemas evaluados.

- Para tener un buen resultado con la red neuronal es necesario tener en cuenta, además de que sea idioma inglés, el lugar del cual es el inglés, es decir, la red neuronal fue entrenada con inglés estadounidense.

- Los fonemas que son difíciles de pronunciar, presentan un nivel más alto de complejidad, eso se pudo evidenciar en el momento de aplicar la red a los fonemas roticos, es por ello que para ello es lo mejor es aplicar algunos cambios a la red obtener mejores resultados en estos.

10. AGRADECIMIENTOS

Quiero expresar mi gratitud a Dios, quien con su bendición llena siempre mi vida y a toda mi familia por estar siempre presentes.

De igual manera mis agradecimientos a los colegas del International Institute of Information Technology in Hyderabad, India. Específicamente al Prof. Anil Kumar Vuppala, que nos brindó acceso a la base de datos de niños con CLP en ingle Hindu, donde gracias a estas colaboraciones se hace posible seguir creciendo en el camino de la investigación.

Finalmente quiero expresar mi más grande y sincero agradecimiento al Dr. Juan Rafael Orozco, y al Dr. Tomas Arias Vergara principales colaborador durante todo este proceso, quienes con su dirección, conocimiento, enseñanza y colaboración permitieron el desarrollo de este trabajo

11.REFERENCIAS

[1] S. Murillo Rendon, J.R. Orozco Arroyave, J.F. Vargas Bonilla, J.D. Arias Londoño3, and C.G. Castellanos Dominguez Automatic Detection of Hypernasality in Children, 2011, 167-174.

[2] J.R. Orozco-Arroyave, S. Murillo-Rendon, A.M. Alvarez-Meza, J.D. Arias-Londoño E. Delgado-Trejos4, J.F. Vargas-Bonilla1 and C.G. Castellanos-Domínguez, Automatic Selection of Acoustic and Non-linear Dynamic Features in Voice Signals for Hypernasality Detection, 2011, 529-532

[3] Ling He; Jing Zhang; Qi Liu; Heng Yin; Margaret Lech, Automatic Evaluation of Hypernasality and Consonant Misarticulation in Cleft Palate Speech.2014, 1298.1301

[4] Ling He, Jing Zhang, Qi Liu, Heng Yin, Margaret Lech & Yunzhi Huang, Automatic Evaluation of Hypernasality Based on a Cleft Palate Speech Database 2015, J Med Syst (2015) pp 39-61

[5] Marzieh Golabbakhsh, Fatemeh Abnavi, Mina Kadkhodaei Elyaderani, David P. Kuehn, Automatic identification of hypernasality in normal and cleft lip and palate patients with acoustic analysis of speech, 2017, J Acoust Soc Am .141-929

[6] Xiyue Wang, Ming Tang, Sen Yang, Heng Yin, Hua Huang & Ling He, Automatic Hypernasality Detection in Cleft Palate Speech Using CNN,2015, pp 3521–3547

[7] Vikram C. Mathad, Kathy Chapman, Julie Liss, Nancy Scherer, and Visar Berisha, Deep Learning Based Prediction of Hypernasality for Clinical Applications, 2020, pp 6554-6558

[8] Henna Raunak Seth Tak, Aarti Pushkar Waknis, Sneha Prakash Kulkarni, Perceptual and instrumental analysis of hypernasality in children with repaired cleft palate 2016 pp: 67-72

[9] Viviane Cristina de Castro Marino, Jeniffer de Cássia Rillo Dutka, Flora Taube Manicardi, Giovana Gifalli 2, Patrick Pedreira Silva, Maria Inês Pegoraro-KrookInfluence of speech stimuli in the auditory perceptual identification of hypernasality in individuals with cleft lip and palate. 2020, 11-32

[10] Vikram C. Mathad, Nancy Scherer, Kathy Chapman, Julie Liss, Visar BerishaAn Attention Model for Hypernasality Prediction in Children, 2021

[11] Vikram Cmathad, Nancy Scherer, Kathy Chapman,Julie Liss,Visar Berisha, A Deep Learning Algorithm for Objective Assessment of Hypernasality in Children with Cleft Palate, 2021, 2986-2996

[12] extracción de características http://bibing.us.es/proyectos/abreproy/12054/fichero/MEMORIA

[13] REDES NEURONALES RECURRENTES: ANÁLISIS DE LOS MODELOS ESPECIALIZADOS EN DATOS SECUENCIALES extraído de https://ucema.edu.ar/publicaciones/download/documentos/797.pdf