Universidad Michoacana de San Nicolás de Hidalgo

Facultad de Ingeniería Eléctrica

División de Estudios de Posgrado

IDENTIFICACIÓN TEXTO-INDEPENDIENTE DE PARLANTES USANDO PULSOS GLOTALES Y REDES CONVOLUCIONALES 1D

TESIS

Que para obtener el grado de

MAESTRO EN CIENCIAS EN INGENIERÍA ELÉCTRICA

presenta

Erick Manuel Ruiz Gaona

Dr. José Antonio Camarena Ibarrola

Director de Tesis

Morelia, Michoacán **Octubre 2022**

IDENTIFICACIÓN TEXTO-INDEPENDIENTE DE PARLANTES USANDO PULSOS GLOTALES Y REDES CONVOLUCIONALES ID

Los Miembros del Jurado de Examen de Grado aprueban la **Tesis de Maestría en Ciencias en Ingeniería Eléctrica** de **Erick Manuel Ruiz Gaona.**

Dr. Jaime Cerda Jacobo
Presidente del Jurado

Dr. José Antonio Camarena Ibarrola
Director de Tesis

Dr. Luis Eduardo Gamboa Guzmán
Vocal

Dr. Felix Calderón Solorio
Vocal

Dr. Luis Valero Elizondo
Revisor Externo

Dr. J. Aurelio Medina Rios
Jefe de la División de Estudios de Posgrado de la Facultad de Ingeniería Eléctrica. UMSNH
(Por reconocimiento de firmas)

UNIVERSIDAD MICHOACANA DE SAN NICOLÁS DE HIDALGO
Agosto 2022

Agradecimientos

A mi asesor de tesis:

Dr. José Antonio Camarena Ibarrola por su guía, comprensión y amistad durante la realización del presente trabajo y por todas sus enseñanzas durante mi trayectoria académica.

A mis profesores:

Dr. Félix Calderón Solorio, Dr. Jaime Cerda Jacobo y Dr. Juan José Flores Romero por su contribución en mi formación académica y personal.

A mi esposa:

María Guadalupe Rincón Solorzano por su apoyo, paciencia y amor incondicional, pero sobre todo por creer en mi e impulsarme a seguir cumpliendo mis sueños.

A mi hijo:

Erick Nicolás Ruiz Rincón por ser el motor que me impulsa a seguir creciendo personal y profesionalmente.

A mi familia:

Mis padres María de la Luz Gaona González, Nicolás Ruiz Lopez y hermano José G. Ruiz Gaona por ser los principales promotores de mis sueños. En mis triunfos y en mis tropiezos, siempre han estado presentes, apoyándome.

A mis compañeros del posgrado:

Por el tiempo y las experiencias que pasamos.

A la Universidad Michoacana de San Nicolás de Hidalgo:

En especial a la División de Estudios de Posgrado de la Facultad de Ingeniería Eléctrica por brindar una educación de calidad y excelencia.

"Sé el cambio que quieres ver en el mundo"

Mahatma Gandhi

Lista de Publicaciones

"Text-Independent Speaker Identification with Glottal Flow and 1D Convolutional Neural Networks"
Antonio Camarena-Ibarrola and Erick Ruiz-Gaona
Enviado a Multimedia Tools and Applications, Springer

Resumen

En la identificación texto-dependiente de parlantes todas las personas que serán registradas por el reconocedor deben pronunciar la misma frase. Sin embargo, en muchas circunstancias tenemos que identificar al parlante independientemente de lo que diga, este es el problema de la identificación texto-independiente de parlantes. Este tipo de identificación es más complicada, ya que no podemos simplemente medir la similitud de una emisión de una palabra o frase con otra emisión realizada por el mismo parlante de la misma palabra o frase, debido a que los sonidos pronunciados no necesariamente son los mismos ni se encuentran en el mismo orden.

En este trabajo proponemos un método para la identificación texto-independiente de parlantes, que consiste en buscar segmentos (marcos) con sonido vocalizado de la señal de voz, estos son los que se emiten mientras vibran las cuerdas vocales, para luego estimar el pulso glotal de cada marco utilizando una técnica de filtrado inverso iterativo de diseño propio. Una vez que se han adquirido los pulsos glotales de todos los marcos con sonido vocalizado junto con la etiqueta del parlante correspondiente, entrenamos una red neuronal convolucional 1D sin capas densas (diferente de la arquitectura convencional). Para identificar a un parlante, también detectamos todos los marcos con sonido vocalizado en su discurso, estimamos el pulso glotal de cada marco, luego usamos la red neuronal convolucional 1D entrenada para identificar al parlante al que pertenece cada marco y usamos un esquema de votación para la decisión final con respecto a la identidad del parlante.

Se utilizan dos bases de datos de elocuciones de voz, una es English Language Speech Database for Speaker Recognition (ELSDSR) de la Universidad Técnica de Dinamarca y la segunda es Acoustic-Phonetic Continuous Speech Corpus (TIMIT), creada por: Massachusetts Institute of Technology, SRI International y Texas Instruments, Inc. Con ambas bases de datos el sistema de identificación texto-independiente de parlantes obtuvo excelentes resultados, superando a la mayoría de trabajos similares. Para el corpus ELSDSR se obtiene una precisión del 100 % y para el corpus TIMIT se obtiene una precisión del 99.5 % con los datos de prueba.

Palabras claves: Identificación texto-independiente de parlantes, Pulso glotal, Filtrado inverso iterativo, Red neuronal convolucional 1D.

Abstract

In text-dependent speaker identification all individuals to be registered by the recognizer are instructed to pronounce the same phrase. However, in many circumstances we have to identify the speaker regardless of what he/she says, this is the text-independent speaker identification problem. This type of identification is more complicated, because we can't just simply measure the similarity of an emission of a word or phrase with another emission made by the same speaker of the same word or phrase, since the sounds pronounced are not necessarily the same or in the same order.

In this work we propose a method for the text-independent identification of speakers, that consists of searching the voiced segments (frames) with vocalized sound of the speech signal, which are the ones emitted while the vocal cords vibrate, then estimate the glottal pulses of each frame using an iterative inverse filtering technique of our own design. Once the glottal pulses from all voiced frames have been collected along with the label of the corresponding speaker, we train a 1D Convolutional Neural Network without dense layers (different from the conventional architecture). For identifying a speaker we also detect all voiced frames in his/her speech, estimate the glottal pulse of each frame, then use the trained 1D convolutional neural network for identifying the Speaker each voiced frame belongs to and use a voting scheme for the final decision regarding the identity of the speaker.

Two databases are used, one is the English Language Speech Database for Speaker Recognition (ELSDSR) from the Technical University of Denmark and the second is the Acoustic-Phonetic Continuous Speech Corpus (TIMIT), created by: Massachusetts Institute of Technology, SRI International and Texas Instruments, Inc. In both databases, the text-independent speaker identification system achieved excellent results, outperforming most similar works. For the ELSDSR corpus, an accuracy of 100 % was achieved and for the TIMIT corpus, an accuracy of 99.52 % was achieved with the test data.

Key words: Text-independent speaker identification, Glottal pulse, Iterative inverse filtering, 1D convolutional neural network.

Contenido

Lista de Figuras

Lista de Tablas

Lista de Símbolos

/a/	Fonema
f	Frecuencia de una señal
f_0	Frecuencia fundamental
T_0	Período fundamental
f_s	Frecuencia de muestreo
T_s	Período de muestreo
$s[n]$	Señal de voz
$u[n]$	Señal del flujo glotal
$h[n]$	Respuesta al impulso del tracto vocal
$r[n]$	Radiación de los labios/fosas nasales
$H(z)$	Filtro del tracto vocal
$U(z)$	Transformada $\mathcal{Z}$ de la excitación acústica a nivel de la glotis
$R(z)$	Filtro de la radiación de los labios/fosas nasales
d	Coeficiente del filtro de radiación
a_p	Coeficientes del filtro del tracto vocal
α_p	Coeficientes del filtro de predicción lineal
R_n	Autocorrelación de tiempo corto
$e[n]$	Error de predicción
b	Sesgo de una neurona artificial
f_a	Función de activación
W	Pesos sinápticos
Y	Salida de una neurona artificial
X	Entrada de una neurona artificial
$x_1(t) * x_2(t)$	Convolución entre las señales $x_1(t)$ y $x_2(t)$

Lista de Acrónimos

ADN: Ácido desoxirribonucleico.

TDSI: Identificación texto-dependiente de parlantes (Text Dependent Speaker Identification, por sus siglas en inglés).

TISI: Identificación texto-independiente de parlantes (Text Independent Speaker Identification, por sus siglas en inglés).

ELSDSR: Base de datos que contiene muestras de voz en idioma inglés, cuyo nombre significa English Language Speech Database for Speaker Recognition.

TIMIT: Es una base de datos que contiene muestras de voz en inglés americano de diferentes sexos y dialectos. El significado de TIMIT es Texas Instruments / Massachusetts Institute of Technology.

CNN: Red neuronal convolucional (Convolutional Neural Network, por sus siglas en inglés).

IA: Inteligencia Artificial.

ANN: Redes neuronales artificiales (Artificial Neural Network, por sus siglas en inglés).

LTU: Unidad de umbral lineal (Linear Threshold Unit, por sus siglas en inglés).

SGD: Descenso de gradiente estocástico (Stochastic Gradient Descent, por sus siglas en inglés).

ReLU: Unidad lineal rectificada (Rectified Linear Unit, por sus siglas en inglés).

MLP: Perceptrón multicapa (Multilayer Perceptron, por sus siglas en inglés).

GMM: Modelos de mezclas de gaussianas (Gaussian Mixture Models, por sus siglas en inglés).

HMM: Modelos ocultos de Marcov (Hidden Markov Models, por sus siglas en inglés).

LPC: Coeficientes de predicción lineal (Linear Prediction Coefficients, por sus siglas en inglés).

MFCC: Coeficientes cepstrales de Mel (Mel Frequency Cepstral Coefficients, por sus siglas en inglés).

WPT: Transformada wavelet (Wavelet Packet Transform, por sus siglas en inglés).

VQ: Cuantificación vectorial (Vector Quantization, por sus siglas en inglés).

CC: Cepstrales complejos (Complex Cepstrum, por sus siglas en inglés).

DCT: Transformada coseno discreta (Discrete Cosine Transform, por sus siglas en inglés).

DFT: Transformada de Fourier discreta (Discrete Fourier Transform, por sus siglas en inglés).

IDFT: Transformada inversa de Fourier discreta (Inverse of the Discrete Fourier Transform, por sus siglas en inglés).

TFF: Teoría Fuente y Filtro.

GCI: Instantes de cierre glótico (Glottal Closing Instant, por sus siglas en inglés).

IAIF: Filtrado inverso iterativo y adaptativo (Iterative Adaptive Inverse Filtering, por sus siglas en inglés).

PSIAIF: Filtrado inverso adaptativo iterativo síncrono del tono (Pitch Synchronous Iterative Adaptive Inverse Filtering, por sus siglas en inglés).

ZZT: Ceros de la transformada $\mathcal{Z}$ (Zeros of the $\mathcal{Z}$-Transform, por sus siglas en inglés).

LF: Modelo glotal de Liljencrants-Fant (Liljencrants-Fant glottal model, por sus siglas en inglés).

LF$^{\mathbf{Rd}}$: Modelo glotal de la Transformada-LF (Transformed-LF glottal model, por sus siglas en inglés).

RMS: Media cuadrática (Root Mean Square, por sus siglas en inglés).

FD-GPE: Estimación del pulso glotal en el dominio de la frecuencia (Frequency Domain - Glottal Pulse Estimate, por sus siglas en inglés).

dB: Unidad de decibel.

GPU: Unidad de procesamiento gráfico (Graphics Processing Unit, por sus siglas en inglés)

FN: Falso negativo (False Negative, por sus siglas en inglés).

FP: Falso positivo (False Positive, por sus siglas en inglés).

FPR: Tasa de falsos positivos (False Positive Rate, por sus siglas en inglés).

PARCOR: Correlación parcial (Partial Correlation, por sus siglas en inglés).

ROC: Características operativas del receptor (Receiver Operating Characteristics, por sus siglas en inglés).

TN: Verdadero negativo (True Negative, por sus siglas en inglés).

TP: Verdaderos positivos (True Positive, por sus siglas en inglés).

TPR: Tasa de verdaderos positivos (True Positive Rate, por sus siglas en inglés).

NTIMIT: Base de datos TIMIT modificada por NYNEX para tener una calidad telefónica.

NYNEX: Antigua compañía de telefonía que participo en la creación de la base de datos NTIMIT. El significado de NYNEX es New York / New England EXchange.

ARPAbet: Conjunto de símbolos de transcripción fonética desarrollados por la agencia de proyectos de investigación avanzada (Advanced Research Projects Agency, por sus siglas en inglés).

IPA: Alfabeto fonético internacional (International Phonetic Alphabet, por sus siglas en inglés).

Capítulo 1

Introducción

La identificación se considera como la serie de características que permiten distinguir a un individuo de los demás [Champod00]. En consecuencia, la identificación de individuos engloba distintos aspectos como lo social, el filosófico, el cultural y el biológico [Saks97]. Debido a la complejidad de este concepto es necesario delimitarlo, en esta tesis utilizamos únicamente el aspecto biológico. Algunos de los rasgos del aspecto biológico para la identificación de individuos son: la textura del iris de los ojos, las huellas dactilares, el ADN, el rostro y la voz. En el caso de este trabajo empleamos como rasgo de identificación la voz.

Es posible identificar y autenticar la identidad de un individuo a través del reconocimiento de los patrones de su voz. Esto es posible porque el aparato fonador de cada ser humano es único. Los rasgos físicos, tanto fonéticos como morfológicos, son particulares a cada persona, lo que los convierte en inmunes a imitaciones. Esta característica le brinda ventaja sobre otros sistemas de identificación. Además, que la identificación de individuos por la voz al ser un micrófono el encargado de capturar la voz, este no tiene que entrar en contacto con la persona y hace posible la identificación a distancia [Cobeta13].

1.1. Planteamiento del problema

La identificación de individuos por la voz también es conocida como reconocimiento de parlantes. En este se extraen los rasgos o características principales de la señal de voz

y estas son comparadas con la colección de características almacenados en el sistema de reconocimiento, y de esta manera el sistema determinará de qué parlante se trata.

El reconocimiento de parlantes tiene diversas aplicaciones como ayudar a resolver situaciones como secuestros virtuales, suplantación de la identidad, fraudes y crímenes, permitir el acceso a lugares privados (físicos y virtuales) y el control de dispositivos electrónicos como hablar con nuestros dispositivos móviles, dictar órdenes a nuestro coche mientras manejamos, y emitir mensajes a sistemas de voces instalados en nuestro hogar. En el ámbito educativo las instituciones pueden utilizar el reconocimiento de voz para ofrecer flexibilidad a los estudiantes con discapacidad visual; por ejemplo, ayudarles a tomar exámenes en línea a través de la autenticación por voz.

Una manera de abordar el problema del reconocimiento de parlantes es por medio de la identificación texto-dependiente de parlantes (Text Dependent Speaker Identification, TDSI por sus siglas en inglés), que se caracteriza porque el parlante que va a ser identificado debe pronunciar una palabra o frase especifica, comúnmente se mide la similitud de la emisión de una palabra o frase con otra emisión realizada por el mismo parlante de la misma palabra o frase, por ejemplo, por medio de la dinámica de la señal de voz.

En esta tesis nos enfocamos en un problema más complejo, la identificación texto-independiente de parlantes (Text Independet Speaker Identification, TISI por sus siglas en inglés), se caracteriza porque no es necesario usar una palabra o frase especifica por el parlante para ser reconocido, sino que basta con que hable y diga cualquier cosa durante unos segundos, inclusive sin importar el idioma en el que hable. En este tipo de identificación ya que no se puede medir solamente la dinámica de la señal de voz, ya que no importa el orden en que aparecen los fonemas de la señal de voz [Castro18].

1.2. Justificación

Desde hace unos años, la sociedad ha normalizado interactuar con los asistentes de los coches, comunicarse con dispositivos móviles y con los asistentes del hogar. La identificación por voz, podría solucionar los problemas con la identificación en nuestro banco o en la factura de la luz. De hecho, servicios como certificar la identidad del interlocutor en

los canales de atención al cliente o responder a las consultas que se realizan a través de la banca móvil o de asistentes (como Alexa, Asistente de Google y Bixby voz) ya figuran entre ellos.

Las personas con discapacidad visual pueden utilizar un sistema de identificación y reconocimiento de voz para ayudar a manejar su problema. Por ejemplo, en lugar de firmar o escribir una contraseña para identificarse, podrían ser identificados por su voz. Las instituciones educativas pueden utilizar el reconocimiento de voz para ayudarles a tomar exámenes en línea a través de la autenticación por voz.

La identificación de parlantes texto-independiente puede aportar en el proceso de identificar a un individuo, al trabajar en conjunto con otros métodos de identificación o de manera independiente cuando las situaciones así lo requieran, como en situaciones en las que se cuenta únicamente con una grabación de audio y es necesario identificar la identidad del individuo de la grabación, así como en un proceso judicial. Además, la TISI es un campo de investigación abierto, ya que se siguen desarrollando nuevas técnicas que mejoren la identificación, por lo que sigue en estudio.

1.3. Hipótesis

Para producir la voz intervienen muchos órganos, músculos, ligamentos, cartílagos y cavidades; estos tienen que estar ajustando continuamente las variables que controlan la conversión de energía aerodinámica en energía acústica. Los pulmones y los músculos por los cuales se comprimen o dilatan estos, crean la fuente de aire que pasa por la glotis y hace vibrar las cuerdas vocales, creando un tren de pulsos glotales, este cuenta con características únicas para cada persona. Posteriormente este tren de pulsos glotales llega al tracto vocal, donde toma la forma de la palabra que se desea producir.

Las redes neuronales convolucionales 2D (CNN, por sus siglas en inglés) han destacado en la visión por computadora, así como en la clasificación de objetos, detección de patologías por medio de radiografías, tomografías, etc. En el reconocimiento de parlantes, las CNN 2D requieren modelar a estos como imágenes, tales como espectrogramas, escalagramas o incluso entropygramas, todas estas imágenes muestran cómo evoluciona la energía

o entropía durante el tiempo, Sin embargo, para el problema de la TISI los parlantes se modelan con independencia del tiempo, no importa el orden en que aparecen los fonemas de la señal de voz. Por lo que usamos la forma de onda del flujo glótico de la voz, que se sabe que varía de una persona a otra independientemente de su ubicación en el tiempo y luego usamos CNN 1D para identificar al parlante.

1.4. Metodología de investigación

Para realizar el análisis del pulso glotal de la señal de voz y la TISI, se utilizan dos bases de datos en el idioma inglés, estas son ampliamente utilizadas en el área de reconocimiento de parlantes por la voz. Estas bases de datos permitirán entrenar a la CNN y posteriormente probar el desempeño del modelo. La primer base de datos utilizada es ELSDSR, este corpus consta de grabaciones de 20 daneses, 1 islandés y 1 canadiense, 10 mujeres y 12 hombres, con edades comprendidas entre los 24 y los 63 años. Cada parlante tiene 9 grabaciones, 7 para entrenamiento y 2 para prueba. La segunda base de datos es TIMIT, este corpus contiene grabaciones de 630 parlantes (438 hombres y 192 mujeres) de ocho dialectos principales del inglés americano, cada uno leyendo 10 oraciones fonéticamente ricas.

La implementación del sistema se lleva a cabo en el lenguaje de programación Python, ya que este es muy versátil y se adapta a diferentes estilos y proyectos. Además, que es un lenguaje que cuenta con una gran cantidad de bibliotecas (en específico de inteligencia artificial, como Keras de Tensor Flow), marcos, extensiones de archivo y colecciones de módulos. Por último, es de código abierto y tiene a sus espaldas a una gran comunidad de desarrolladores que siempre están mejorándolo y perfeccionándolo.

El sistema comienza dividiendo la señal de voz en segmentos (marcos) de 30 ms y revisando cual de estos contienen sonido vocalizado, debido a que el pulso glotal solo está presente en este tipo de sonidos. Se estima el pulso glotal por medio de un método propio, este es similar a las técnicas de Murphy [Murphy08] o de Alku [Alku92a]. Estos métodos se basan en la teoría fuente-filtro de Fant [Fant60] y se conocen como métodos de filtrado inverso.

Se diseña un modelo CNN 1D para la clasificación o identificación de los parlantes, la CNN contiene muchos hiperparámetros y variables para su diseño, como el número de capas, tamaño de kernel, función de activación, tasa de aprendizaje, optimizador, regularizadores, etc., lo que las vuelve complejas de diseñar. Una vez seleccionado y entrenado el modelo CNN 1D con la colección de pulsos glotales etiquetados con el parlante que emitió el sonido, se utiliza para la identificación de parlantes. En la Figura 1.1 se muestra un resumen de los principales pasos utilizados en esta tesis para la identificación de parlantes.

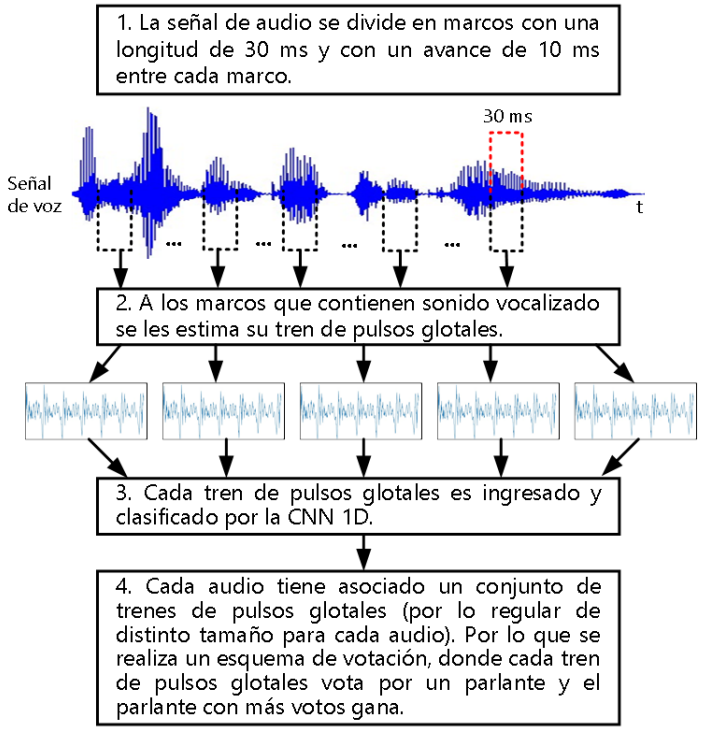

Figura 1.1: Proceso de identificación de parlantes por medio del pulso glotal

1.5. Antecedentes

La identificación de parlantes se ha estudiado desde hace décadas, dando como resultado una mayor comprensión de la producción de voz y los factores que hacen a la voz humana un rasgo de identificación. Los trabajos publicados en esta área se centran sus esfuerzos en la extracción de características, así como en los métodos que se pueden emplear

para comparar las características.

La extracción de características ha sido abordada desde distintas perspectivas, ya sea tratando de modelar el tracto vocal, la fuente de excitación, o utilizando el conocimiento acerca del oído humano para obtener información del espectro de frecuencia de la señal de voz. Entre las técnicas que han tenido una mayor relevancia se encuentran la codificación lineal predictiva (LPC, por sus siglas en inglés), los coeficientes cepstrales de Mel (MFCC, por sus siglas en inglés), transformada wavelet (WPT, por sus siglas en inglés) y la transformada de Fourier discreta (DFT, por sus siglas en inglés).

En el proceso de comparación de características, se toman las características extraídas de la señal de voz del parlante a identificar y se busca encontrar similitudes con las características utilizadas para crear las plantillas de la base de datos del sistema. Las soluciones más simples utilizan las distancias para comparar características, donde las distancias más comúnmente utilizadas se encuentra la distancia Manhattan y distancia Euclidiana, además de la medida de divergencia coseno. Con el paso del tiempo se abordó el problema de la comparación de características utilizando técnicas más complejas como la cuantificación vectorial (VQ, por sus siglas en inglés), mezcla de gaussianas (GMM, por sus siglas en inglés), modelos ocultos de Marcov (HMM, por sus siglas en inglés) y redes neuronales artificiales (ANN, por sus siglas en inglés).

La identificación de parlantes es un área en continuo crecimiento, ya que se siguen desarrollando nuevas técnicas de extracción de características de la señal de voz y utilizando nuevos modelos para su clasificación. En esta tesis para la extracción de características se propone un método propio que estima el pulso glotal de la señal de voz, por medio de la técnica LPC. Para el proceso de comparación de características utilizamos CNN 1D. Es importante conocer los trabajos más exitosos en el reconocimiento de parlantes para tener un punto de partida y comparación.

- Plumpe en 1999 [Plumpe99], utilizó la derivada del flujo glotal como características de la voz. La estimación de la derivada del flujo glótico la descompone en una estructura gruesa, que representa la forma general del flujo, y una estructura fina, que comprende la aspiración y otras perturbaciones en el flujo, a partir de las cuales se obtienen los

parámetros del modelo. La derivada del flujo glótico la estima utilizando un filtro inverso determinado dentro de un intervalo de tiempo de cierre de las cuerdas vocales que se identifica a través de las diferencias en la modulación de la frecuencia de los formantes durante las fases abierta y cerrada del ciclo glótico. La estimación de la derivada del flujo glótico se modela utilizando el modelo de Liljencrants-Fant [Fant85] para capturar su estructura gruesa, mientras que la estructura fina de la derivada del flujo se representa a través de medidas de energía y perturbación. Además, utiliza GMM para la clasificación de los parlantes. Para la base de datos TIMIT logra en promedio una identificación correcta de un 70 %.

- Soong en 1987 [Soong85], utilizó VQ para la clasificación de los parlantes, con una base de datos de grabaciones telefónica de 100 parlantes (50 hombres y 50 mujeres) que constaba de expresiones de dígitos aislados. Para diez dígitos aislados aleatorios pero diferentes, se logró una precisión de identificación de parlantes superior al 98 %.

- Reynolds en 1995 [Reynolds95], utilizó MFCC y para la clasificación utiliza GMM. Obtiene excelentes resultados con las bases de datos TIMIT y NTIMIT (NTIMIT es la versión con calidad telefónica de TIMIT) con una exactitud de identificación de 99.5 y 60.7 %, respectivamente. Sin duda, el sistema de Reynolds ha sido un referente para la identificación de parlantes, y por esta razón muchos trabajos han realizado modificaciones para intentar mejorarlo [Banerjee18, Al-Rawahy12, Guonason08, Thyes00, Nakagawa12, Veena15].

- Rawahy en 2010 [Al-Rawahy10], implementa un sistema de identificación texto-independiente de parlantes basado en la transformada Zak. Utiliza la base de datos ELSDSR, obteniendo una eficacia de identificación de 100 % con los archivos de prueba.

- Saady en 2014 [Saady14], presenta un sistema de reconocimiento de parlantes mediante el uso de WPT y ANN. Utiliza la base de datos ELSDSR, obteniendo una eficacia de identificación de 95.7 % con los archivos de prueba.

- Castro en 2018 [Castro18], implementa un sistema de reconocimiento de parlantes utilizando los tres primeros formantes y creando con estos una nube de puntos tridi-

mensional. Con la base de datos ELSDSR obtuvo una exactitud del 90 %.

- Reynoso en 2021 [Reynoso21], utilizó los formantes de la voz para generar imágenes y para la clasificación utiliza CNN 2D. Se utilizó una base de datos que cuenta con 2,856 archivos de audio. La base de datos fue creada utilizando a 21 personas, 6 mujeres y 15 hombres. Cada persona pronunció 34 palabras los números del cero al nueve y las letras del alfabeto griego. Además, cada una de las palabras es pronunciada cuatro veces. Se obtuvo una exactitud del 96 %.

1.6. Objetivos de la Tesis

1.6.1. Objetivos generales

Diseñar un sistema que reconozca a personas por medio de su señal de voz, independiente de las palabras contenidas en dicha señal. El sistema estima el pulso glotal de la señal de voz del parlante y este es introducido a una red neuronal convolucional para ser clasificado.

1.6.2. Objetivos específicos

1. Diseñar un clasificador de voz que sea capaz de distinguir segmentos de audio con contenido de sonidos vocalizados y no vocalizados.
2. Diseñar un extractor de pulsos glotales a partir de segmentos de audio con contenido de sonidos vocalizados.
3. Entrenar una red neuronal convolucional 1D con la colección de pulsos glotales etiquetados con el parlante que emitió el sonido.
4. Utilizar la red convolucional ya entrenada para identificar al parlante.

1.7. Descripción de capítulos

- Capítulo 2, describe como se produce la voz, así como sus características. Se explica la teoría fuente-filtro y la extracción de características por medio del análisis de pre-

dicción lineal. Además se presenta el método utilizado para la estimación del pulso glotal.

- Capítulo 3, se presenta una breve introducción a las redes neuronales (en especial las CNN), como conceptos básicos, tipos de redes neuronales, hiperparámetros y soluciones a los problemas que pueden presentan las redes neuronales. Además de la arquitectura de la CNN utilizada.

- Capítulo 4, detalla el procedimiento para la estimación del pulso glotal y la descripción de las bases de datos utilizadas.

- Capítulo 5, presenta los experimentos realizados y los resultados obtenidos para cada base de datos.

- Capítulo 6, expone las conclusiones finales del trabajo realizado en esta tesis y el trabajo futuro.

Capítulo 2

Estimación del Pulso Glotal

2.1. Producción de la voz

Para realizar un reconocimiento texto-independiente de parlantes, es esencial entender cómo los humanos producen las señales de la voz, por lo tanto, se dará una descripción general de los mecanismos de generación de sonido en la producción de la voz humana. Una discusión exhaustiva de fonética y lingüística está fuera del alcance de este trabajo, sin embargo, existe una amplia variedad de libros [Cobeta13, Flanagan72, Stevens98] que dan detalles sobre los temas de fonética y lingüística.

2.1.1. Fisiología de la producción de la voz

La producción de la voz o fonación es una función sobreañadida, esta es fruto de la evolución humana a pesar de que no se trata de una función esencial para sobrevivir, ya que, se trata de una actividad generada por las personas para relacionarse. Las dos funciones biológicamente primarias de la laringe son: la respiratoria y la esfinteriana [Guitart04].

Según Obediente [Obediente98], durante la fonación se produce un continuo ajuste del flujo aéreo por la interacción de las estructuras subglóticas, glóticas y supraglóticas, modificando las variables que controlan la conversión de la energía aerodinámica en energía acústica: la presión subglótica, las propiedades biomecánicas de las cuerdas vocales y la resistencia supraglótica, y la resonancia. En la Figura 2.1 se muestran las principales es-

tructuras anatómicas involucradas en el proceso de fonación. Por lo general, se acostumbra dividir el aparato fonador en tres grandes grupos:

- Cavidades infraglóticas o subglóticas
- Cavidad laríngea o glótica
- Cavidades supraglóticas

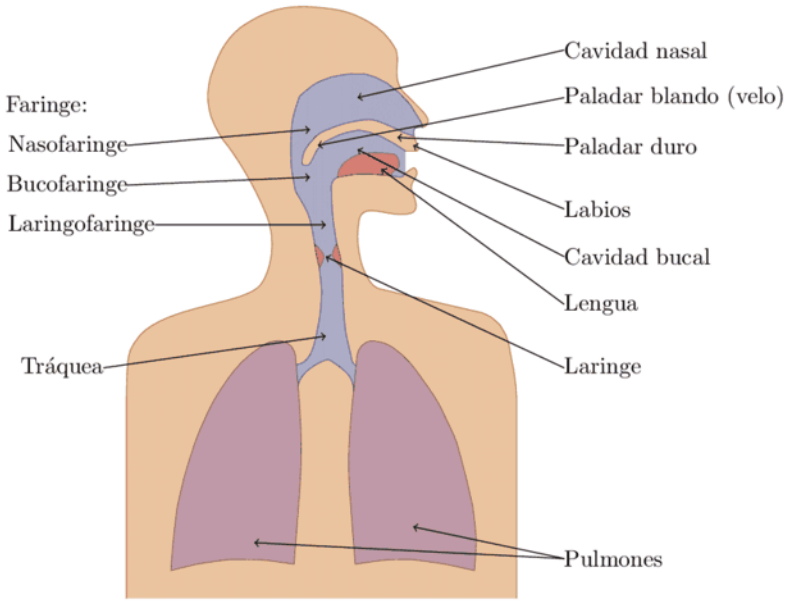

Figura 2.1: Esquema representativo del aparato fonador. Se indican las principales estructuras anatómicas involucradas en el proceso de fonación. Fuente [Alzamendi16].

Cavidades subglóticas

De acuerdo a Cobeta [Cobeta13], las cavidades subglóticas o infraglóticas son la fuente de energía para todo el proceso de producción de voz, aunque su función principal es la respiración, asegurando el abastecimiento de oxígeno a la sangre. En el proceso de inspiración, los pulmones toman aire, baja el diafragma y agranda la cavidad torácica. En el momento de la fonación, la espiración, provocada por la contracción de los músculos intercostales y del diafragma, aporta la energía necesaria para generar la onda de presión acústica que atravesará los órganos fonadores superiores.

La presión subglótica es la energía aerodinámica de entrada al aparato fonador y se genera en las vías respiratorias bajas. Durante la espiración se establece un flujo aéreo desde los pulmones hacia la glotis, gracias a que la presión intratorácica excede la atmosférica. Durante el habla, las dimensiones y la forma de las vías respiratorias altas y de la propia laringe cambian constantemente, afectando a la presión subglótica. Los cambios en la geometría de la glotis y las propiedades viscoelásticas de las cuerdas que se asocian a los movimientos articulatorios de éstas pueden alterar el umbral de la presión subglótica necesaria para mantener la vibración vocal, lo que se conoce como "presión umbral de fonación". Se cree que existen sensores de presión, propioceptivos, de tensión y estiramiento muscular, y auditivos, que ayudan a controlar la presión subglótica necesaria para comenzar y mantener la producción de la voz [Cobeta13].

Cavidad laríngea o glótica

La laringe, también llamada órgano vocal, es un sistema complejo formado por cartílagos, músculos y ligamentos. La laringe es el órgano responsable de la producción básica de la voz, se sitúa en la parte medial y anterior del cuello, por delante de la faringe. Cranealmente comunica, a través de la faringe, con la cavidad bucal y las fosas nasales, y caudalmente se continúa con la tráquea. Interviene en la respiración, la deglución y la fonación, su función principal es la de actuar como válvula de cierre de la vía que conecta hacia los pulmones, evitando que pasen los alimentos a través de la tráquea [Cobeta13]. En la Figura 2.2 se muestran los principales cartílagos, ligamentos y membranas que componen la laringe.

El último cartílago de la tráquea, el cricoides, forma la base de la laringe. El principal órgano de la laringe son las cuerdas vocales que son dos pares de pliegues compuestos de ligamentos y músculos. El par inferior son las llamadas cuerdas vocales verdaderas o pliegues vocales y las llamadas cuerdas vocales falsas son los pliegues vestibulares. La laringe esta protegida en su parte anterior por el cartílago tiroides, abierto por su parte posterior. Finalmente, la parte superior de la laringe está unida al hueso hioides [Cobeta13]. En la Figura 2.3 se muestra una imagen generada con un videoestroboscopio de las cuerdas vocales.

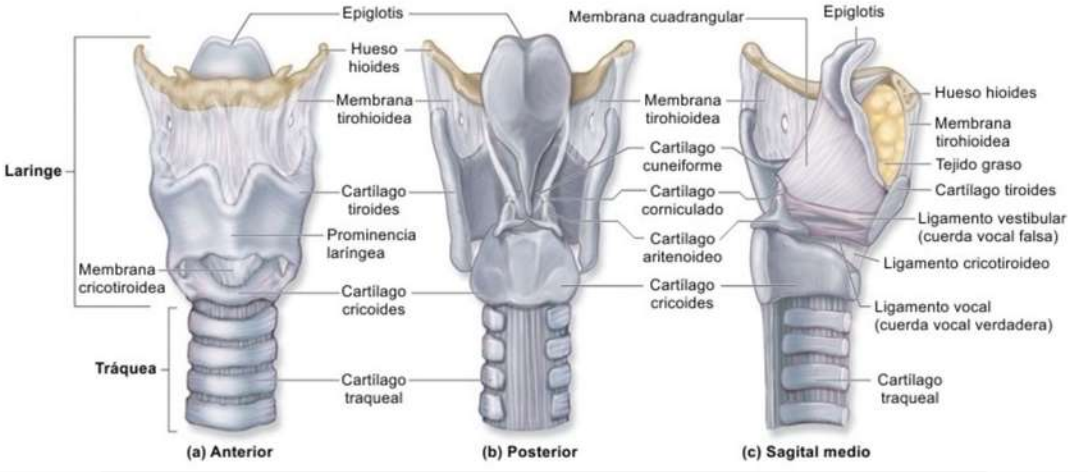

Figura 2.2: Anatomía de la laringe, expuesta a partir de una vista anterior (a), una vista posterior (b) y un corte sagital medio (c). Pueden observarse los principales cartílagos, ligamentos y membranas que componen la laringe. Se muestran también las estructuras anatómicas que se comunican directamente con la laringe. Fuente [Alzamendi16].

Las cuerdas vocales se ubican dentro de la laringe, en la parte superior de la tráquea que se une posteriormente a los cartílagos aritenoides, y de manera anterior al cartílago tiroides. Los músculos responsables del movimiento de estas son: músculo cricotiroideo, músculo cricoaritenoideo posterior, músculo cricoaritenoideo lateral, músculo tiroaritenoideo y músculo aritenoideo. Gracias a estos músculos las cuerdas vocales puedan juntarse (aducción) o separarse (abducción). La abertura entre ambas cuerdas se denomina glotis, esta cuenta con una parte cartilaginosa y otra ligamentosa. Existen diferencias

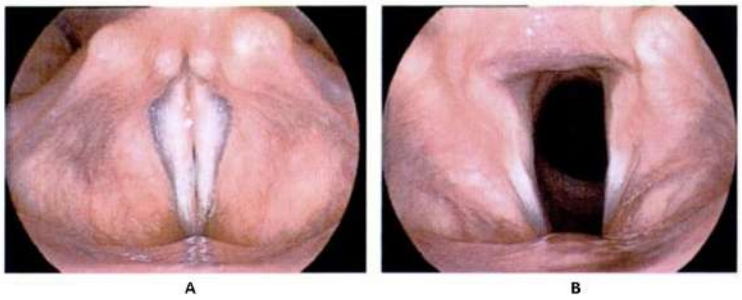

Figura 2.3: Vista de las cuerdas vocales, generada con un videoestroboscopio. A, cuerdas vocales en aducción; B, cuerdas vocales en posición de respiración. Fuente [A.02].

según el sexo en las dimensiones absolutas y relativas de la glotis; las diferencias de longitud son estadísticamente significativas y son la causa de la diferente frecuencia fundamental del hombre y la mujer [Cobeta13]. La Figura 2.4 muestra la configuración de las cuerdas vocales en la respiración, y en la generación de sonidos sonoros, y sordos.

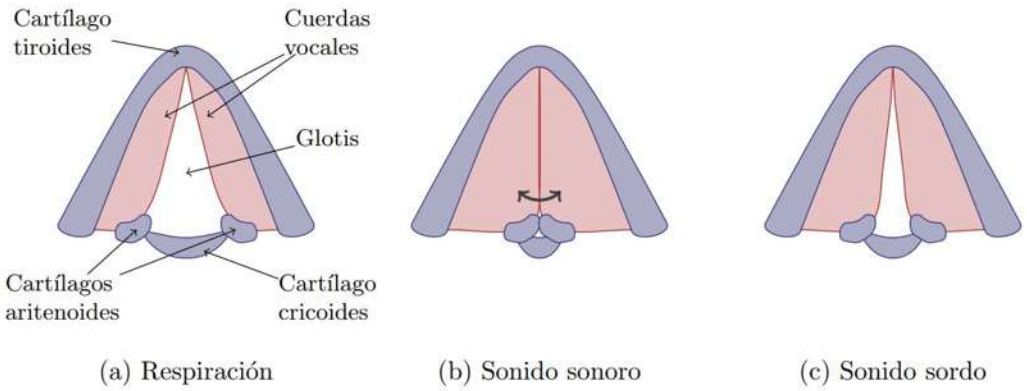

Figura 2.4: Configuración de las cuerdas vocales en la respiración (a), y en la generación de sonidos sonoros (b), y sordos (c). Cada esquema representa un corte horizontal de la laringe, a la altura de las cuerdas vocales. Fuente [Alzamendi16].

Según Miyara [Miyara22b], a través de la glotis pasa el aire que permite la vibración de las cuerdas vocales y, por lo tanto, la generación básica de la voz. Si la glotis se cierra por completo, el sonido no se produce. La glotis también posibilita que el aire llegue a los pulmones. Por encima de la glotis se halla la epiglotis, una estructura cartilaginosa que bloquea la glotis a la hora de la deglución: de este modo, la comida no se introduce en el sistema respiratorio. El sonido producido por las cuerdas vocales consta de una frecuencia fundamental y sus armónicos superiores. El tono de la voz está relacionado con la longitud y grosor de las cuerdas vocales de cada individuo. Las fases principales de la glotis se muestran en la Figura 2.5 y se describen a continuación:

1. Cuando las cuerdas vocales se encuentran separadas, la glotis adopta una forma triangular. El aire pasa libremente y prácticamente no se produce sonido, es el caso de la respiración.

2. Cuando la glotis comienza a cerrarse, el aire que la atraviesa proveniente de los pulmones experimenta una turbulencia, emitiéndose un ruido de origen aerodinámico

conocido como aspiración (aunque en realidad acompaña a una espiración o exhalación).

3. Al cerrarse más la glotis, las cuerdas vocales comienzan a vibrar a modo de lengüetas, produciéndose pulsos cuasi-periódicos (conocidos como pulsos glotales).

4. Finalmente, es posible obturar la glotis completamente. En ese caso no se produce sonido.

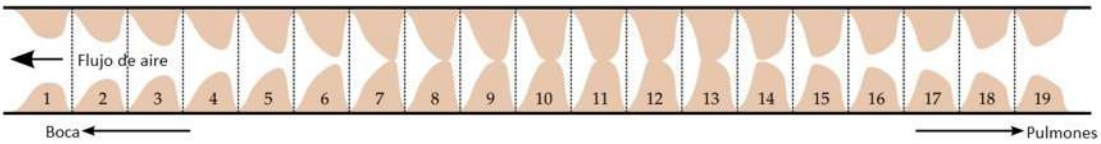

Figura 2.5: Serie temporal esquemática de la vibración de las cuerdas vocales. Los fotogramas 1 a 6 ilustran la fase de cierre. La glotis se cierra completamente durante los fotogramas 7 a 13, durante los cuales el área de contacto se mueve hacia arriba, hasta que los pliegues se separan nuevamente en los fotogramas 14 a 19. A lo largo del ciclo, la parte inferior de las cuerdas vocales (en el lado derecho de la ilustración) lidera el movimiento. Fuente [Airas08].

Cavidades supraglóticas

Todas las cavidades situadas por encima de los pliegues vocales (cuerdas vocales) actúan, o pueden actuar, como cajas de resonancia de la voz. Se habla de resonadores o cavidades supraglóticas [Torres07]. La faringe es la cavidad situada entre la epiglotis y la úvula; las paredes musculares que la circunscriben pueden modificar su volumen y su forma [Obediente98].

De acuerdo a Alzamendi [Alzamendi16], las vías aéreas supraglóticas están formadas, principalmente, por la faringe y las cavidades oral y nasal. La faringe, o garganta, es un conducto en forma de embudo, cuyas principales funciones son permitir el pasaje de aire, en la respiración y en la fonación, y de líquidos y alimentos, en la deglución. Se divide en tres regiones (ver Figura 2.1):

- Laringofaringe: región inferior ubicada por encima y por detrás de la laringe;
- Bucofaringe u orofaringe: porción central localizada por detrás de la cavidad bucal;

- Nasofaringe: región superior ubicada por detrás de la cavidad nasal.

Las cavidades supraglóticas le imprimen al tono laríngeo las características que darán como resultado el timbre definitivo del sonido. La corriente de aire que proviene de la glotis sufre en estas partes una serie de modificaciones debidas a contactos y estrechamientos que se producen en determinados puntos de su trayectoria; estas actúan, a manera de filtros que solo dejan pasar las frecuencias que coinciden con las de estas cavidades de resonancia. La forma, tamaño y volumen de estas cavidades determinan que armónicos conservarán, disminuirán o aumentarán su energía, es decir, la composición espectral de la onda sonora. El tracto vocal posee cinco resonancias importantes, llamados formantes, que definen la calidad y tipo de vocal cuando es audible, los órganos articulatorios del tracto vocal influyen en la emisión de los sonidos vocálicos, ya que cada sonido vocálico tiene distinta forma de pronunciación en el tracto vocal [Obediente98].

2.1.2. Pulso Glotal

De acuerdo a las Teorías mucoondulatoria de Perelló (1962) [Perelló62] y mioelástica-aerodinámica de Van den Berg (1958) [denBerg58], las cuerdas vocales se aproximan, se contraen y se tensan durante la fonación para regular su elasticidad. Además de regular la tensión vocal y la elasticidad, el control neuromuscular también ajusta la configuración de la apertura glótica. El perfil dinámico tridimensional de la glotis determina la diferencia entre las presiones subglótica y supraglótica, con lo cual la configuración de la apertura glótica es un componente importante de la fuerza aerodinámica motora de la fonación. Así, para iniciar la voz, las cuerdas vocales deben aproximarse para formar un canal estrecho o ligeramente cerrado que separa la subglotis de la supraglotis. Una vez que la glotis está cerrada o casi cerrada, comienza la espiración de aire desde los pulmones, con lo que aumenta la presión entre las cuerdas y se produce un empuje en contra de su elasticidad. Cuando la presión del aire es lo bastante alta como para poder separar los tejidos de las cuerdas (estando los aritenoides unidos), el aire fluye a través de la apertura glótica generada. La diferencia entre la presión subglótica y la supraglótica (atmosférica) produce una presión positiva que insufla aire desde la tráquea hacia la superficie medial de las

cuerdas vocales. En cuanto el flujo aéreo pasa a través del estrechamiento del conducto que determina la glotis, la velocidad de sus moléculas aumenta, determinando una reducción de la presión transglótica (la presión transglótica se define como la diferencia entre la presión subglótica y la presión supraglótica) que produce una presión negativa. Una vez que el aire fluye por la ahora abierta glotis, varias fuerzas se combinan para cerrarla de nuevo. Hay tres fuerzas principales que intervienen en el cierre de la glotis: el efecto Bernoulli del flujo aéreo a través de un estrechamiento del conducto crea una fuerza negativa que tracciona de la cuerda medialmente; la elasticidad o retroceso pasivo de las cuerdas vocales hace que éstas recobren su forma original antes de haber sido deformadas por la presión transglótica; y el aire escapando a través de la glotis desde la región subglótica hace que caiga la presión subglótica y descienda la fuerza que mantiene apartados los tejidos de las cuerdas vocales. Todos estos factores llevan a que las cuerdas se cierren hacia su posición de aproximación, para obstruir nuevamente el flujo aéreo e incrementar otra vez la presión subglótica hasta que pueda deformar los tejidos de las cuerdas e iniciar otro ciclo de la fase abierta. Este ciclo de vibración se denomina "ciclo glótico".

El flujo glótico es la forma de onda de la velocidad del flujo de aire que sale de la glotis y entra en el tracto vocal, también llamado fuente glótica. La forma de onda de la fuente glótica en un solo período se indica mediante el pulso glótico. Cabe señalar que es muy común encontrar en varios textos de investigación el uso de los términos flujo glótico, fuente glótica y pulso glotal indistintamente. Sin embargo, Murphy [Murphy08] afirma que el flujo glótico y el pulso glotal representan el flujo de aire que pasa a través de la glotis y que la fuente glótica es la onda de presión glótica o, en algunos casos, la apertura glótica, es decir, la derivada del pulso glotal.

Cuando las cuerdas vocales abren y cierran la glotis a intervalos idénticos, la frecuencia del sonido generado es igual a la frecuencia de vibración de las cuerdas vocales [Sundberg87]. Cada ciclo consta de cuatro fases glóticas, como se puede observar en la Figura 2.6: cerrado, abriendo, abierto y cerrando (o retorno).

Por lo general, cuando las cuerdas vocales están en una posición cerrada, el flujo comienza lentamente, se acumula hasta un máximo y luego disminuye rápidamente a cero cuando las cuerdas vocales se cierran abruptamente [kafentzis08]. Sin embargo, los estudios

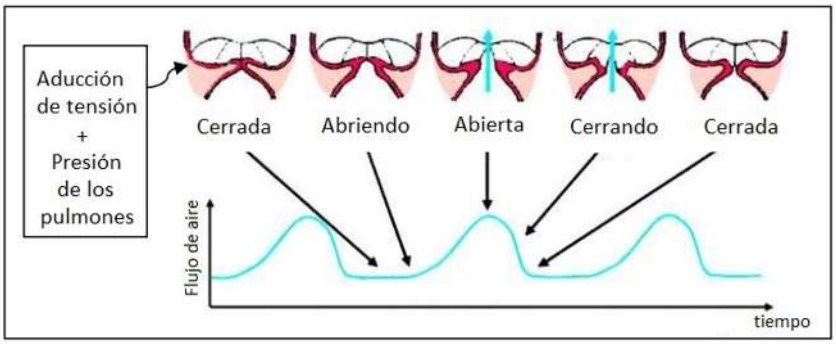

Figura 2.6: Las fases glóticas. Fuente [deOliveira Dias12].

han indicado que el cierre total de la glotis es una suposición idealista y que la mayoría de los individuos presentan algún tipo de fuga glótica durante la supuesta fase de cierre. Aún así, la mayoría de los modelos glóticos asumen que la fuente tiene flujo cero durante la fase cerrada del ciclo glótico [deOliveira Dias12].

El intervalo de tiempo durante el cual las cuerdas vocales están cerradas y no se produce flujo se denomina fase de cierre glótico. La siguiente fase, durante la cual hay un flujo distinto de cero y hasta el máximo de la velocidad del flujo de aire, se denomina fase de apertura glótica, y el intervalo de tiempo desde el máximo hasta el momento del cierre glótico se denomina fase de cierre o retorno. Muchos factores pueden influir en la velocidad a la que las cuerdas vocales oscilan a través de un ciclo cerrado, abierto y de retorno, como la tensión de los músculos de las cuerdas vocales, la masa de las cuerdas vocales y la presión del aire debajo de la glotis [kafentzis08].

Durante la fase abierta de la glotis, el tracto vocal influye en la producción del pulso, ya que actúa como una carga para el tracto glotal. Esta influencia tiene varios efectos, de los cuales se presentan a continuación los dos más importantes [Duxans00]:

- Desviación (skewness): desplazamiento del área del pulso glotal hacia la derecha. Su efecto acústico es un incremento uniforme del nivel de potencia de los formantes.

- Rizado: se puede detectar un rizado en la fase abierta del pulso debido a la disipación de energía del primer formante por parte de la glotis. Esto puede provocar un desplazamiento en las frecuencias de los formantes y un incremento de su ancho de banda

durante la fase abierta del ciclo glotal. Acústicamente su efecto es una reducción en el nivel del primer formante.

Debido a esta interacción, la forma del pulso glotal varía ligeramente cuando hay un cambio en el tracto vocal. Es decir, el pulso glotal presenta cierta dependencia respecto al contenido fonético de la voz producida. Por lo tanto, la estimación del pulso debe realizarse para cada período de tono (pitch), o como máximo para cada tramo de la señal con propiedades fonéticas estables [Duxans00].

2.1.3. Percepción del sonido

La audición, en su definición más simple, es la capacidad de los humanos y los animales para percibir el sonido. El oído es el sensor que permite que el ser humano escuche el sonido. La percepción del sonido implica hacer interpretaciones de la naturaleza de la señal sonora percibida, esta es el resultado del procesamiento neuronal que ocurre principalmente después de que el oído detecta el sonido y analiza su espectro [Rabiner11].

Las ondas sonoras se pueden caracterizar por su amplitud y frecuencia de variación. Estas son características que se pueden representar matemáticamente y medir utilizando dispositivos físicos. Los seres humanos, sin embargo, sienten estas características del sonido y las perciben como volumen y tono respectivamente, y estas características perceptivas están relacionadas con la amplitud y la frecuencia [Rabiner11]. Las ondas sonoras tienen una frecuencia entre 20 Hz y $20,000$ Hz. Las ondas acústicas de menos de 20 Hz se denominan infrasonidos, y los de más de $20,000$ Hz se llaman ultrasonidos. Por lo general ni unos ni otros son audibles por el ser humano. La voz se ubica en un rango de frecuencias aproximadamente entre $50-8,000$ Hz, sin embargo se puede representar de una manera aceptable en un rango de $50-3,500$ Hz [Miyara22a].

Frecuencia fundamental, tono y timbre de la voz

La frecuencia fundamental (f_0) representa el número de veces que las cuerdas vocales se abren y cierran por segundo, se expresa en ciclos por segundo o Hz. La laringe humana es capaz de producir una amplia gama de frecuencias (rango vocal), que varía en

función de la edad y del sexo. Los valores normales son de unos 125 Hz para el hombre, 250 Hz para la mujer y 350 Hz en la infancia. La f_0 puede variar dentro de unos límites determinados en función de: la masa de las cuerdas vocales, la longitud y la tensión de las cuerdas vocales y la presión subglótica [Cobeta13].

La percepción psicoacústica del hecho físico de la frecuencia es el tono vocal. El tono (también llamado pitch) percibido no depende únicamente de la f_0, ya que otros parámetros, como la intensidad o la composición espectral también desempeñan un papel, aunque secundario, cuando aumenta la f_0 el tono se hace más agudo, y cuando disminuye se hace más grave. Estos cambios no son lineales y no percibimos igual el mismo aumento a una frecuencia baja que a una frecuencia alta. Por ejemplo, el paso de 100 a 150 Hz es más evidente para nuestros oídos que el de $2,500$ a $2,550$ Hz. Las notas musicales reflejan este fenómeno de percepción, y así, el paso del "do"de la primera octava al "do"de la segunda es de 32.7 a 65.4 Hz, mientras que el paso del "do"de la quinta al "do"de la sexta octava es de 523.2 a $1,046.5$ Hz: para subir una octava hay que duplicar la frecuencia en el rango de las frecuencias altas [Cobeta13].

De acuerdo a Rabiner [Rabiner11], así como el volumen (subjetivo) de un tono está influenciado tanto por la intensidad del sonido como por la frecuencia del sonido, el tono percibido está influenciado de manera similar tanto por la frecuencia del sonido como por la intensidad del sonido. En general, el tono (el atributo subjetivo) está altamente correlacionado con la frecuencia (el atributo físico). La unidad de frecuencia es Hz y la unidad de tono es el mel (derivado de la palabra melodía). En la Figura 2.7 se muestra que la relación entre el tono y la frecuencia de un tono puro no es lineal [Rabiner11]. La curva se ajusta con precisión a la siguiente ecuación:

$$\text{Tono en mels} = 1127 \log_e(1 + f/700),$$

donde, f es la frecuencia del tono.

El timbre es la propiedad de la voz que nos permite distinguir entre dos notas de igual frecuencia e intensidad emitidas por instrumentos musicales distintos, o diferenciar

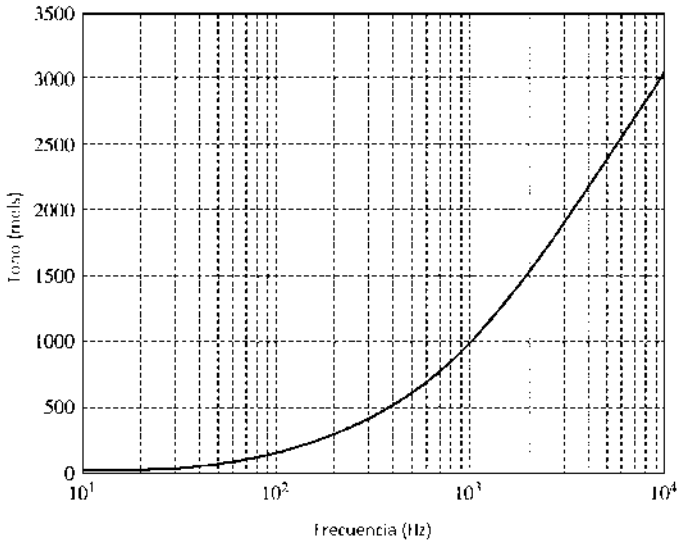

Figura 2.7: Gráfico del tono subjetivo (en mels) frente a la frecuencia real (en Hz) de un tono puro. Fuente [Rabiner11].

dos voces pertenecientes a personas distintas. El timbre depende de los formantes y de las dimensiones físicas del tracto vocal, de la frecuencia fundamental y de la intensidad. Si se alteran las amplitudes relativas de los armónicos de un sonido y sus fases con relación al tono fundamental, varía el timbre del sonido sin cambiar su tono [Cobeta13].

Intensidad y volumen de la voz

La intensidad se define como la amplitud de la variación de la presión sonora producida al transmitirse la voz en el medio aéreo, y se expresa en decibelios (dB). Para un adulto normal, la intensidad de la fonación durante la conversación está entre 75 y 80 dB. El valor de la intensidad depende fundamentalmente de la amplitud de la vibración de las cuerdas vocales y de la presión subglótica. La intensidad vocal es un importante factor en la comunicación y se encuentra regulado en los tres niveles: subglótico, glótico y supraglótico [Cobeta13].

La sensación psicoacústica del fenómeno físico de la intensidad es el volumen. El volumen relativo de la voz puede determinarse como el valor de la presión sonora (intensidad) de la señal acústica medida en la boca. El volumen percibido, que es subjetivo y solo se puede

determinar a través de experimentos psicofísicos, está influenciado tanto por la intensidad del sonido como por la frecuencia del sonido[Rabiner11].

2.1.4. Tipos de sonidos de la voz y fonemas

La señal de voz tiene las siguientes propiedades inherentes [Rabiner11]:

- La voz es una secuencia de sonidos en constante cambio.
- Las propiedades de la forma de onda de la señal del habla dependen en gran medida de los sonidos que se producen para codificar el contenido del mensaje implícito.
- Las propiedades de la señal de voz dependen en gran medida del contexto en el que se producen los sonidos; es decir, los sonidos que ocurren antes y después del sonido actual. Este efecto se denomina coarticulación del sonido de la voz y es el resultado del mecanismo de control vocal, que anticipa los siguientes sonidos mientras produce el sonido actual, modificando así las propiedades del sonido actual.
- El estado de las cuerdas vocales y las posiciones, formas y tamaños de los diversos articuladores (labios, dientes, lengua, mandíbula, velo, etc) cambian lentamente con el tiempo, produciendo así los sonidos del habla deseados.

De acuerdo a la lista anterior de propiedades, es posible determinar algunas de las propiedades físicas de la voz (si las cuerdas vocales están vibrando o en una posición laxa, si el sonido es cuasi-periódico o similar al ruido, etc.).

La transcripción fonética (o notación fonética) es un sistema de símbolos gráficos para representar los sonidos de la voz humana. Una representación fonética de las palabras de los sonidos que se deben hacer para "pronunciar"las palabras correctamente en un idioma dado. El hecho de que un sonido esté caracterizado por el tipo de excitación y la configuración del tracto vocal nos lleva a definir las unidades lingüísticas básicas del habla, llamadas "fonemas". En realidad, los fonemas son modelos de los sonidos que pueden diferir luego en su expresión acústica, dando lugar a lo que se conoce como "alófonos", se les puede definir como el conjunto mínimo de unidades que permite construir cualquier palabra en un idioma determinado. Así pues, a grosso modo, dos fonemas son distintos si el cambio

de uno por otro cambia la palabra [Cobeta13]. Un mismo fonema puede tener asociado un conjunto de alófonos y su ocurrencia dependerá del contexto en que se encuentre el fonema dentro de una palabra [Rabiner11].

Las palabras están compuestas por sílabas y a su vez, cada sílaba esta compuesta por uno o varios fonemas y pueden ser vocálicos o consonánticos. Una vocal sola puede conformar una sílaba, como en los casos de azul (a - zul) y único (ú - ni - co). Sin embargo, las consonantes siempre necesitan estar acompañadas de vocales para convertirse en sílabas.

Según Rabiner [Rabiner11], en particular, para el inglés americano, hay entre 39 y 48 fonemas, que incluyen vocales, diptongos, semivocales y consonantes. La Figura 2.8 proporciona un conjunto reducido de 39 fonemas del inglés americano junto con la representación ARPAbet [diseñada por la Agencia de Proyectos de Investigación Avanzada (ARPA) para la transcripción en teclados de computadora sin fuentes especiales]. Los símbolos ARPAbet se usan generalmente en ingeniería e informática. Y los símbolos IPA se usan más comúnmente en publicaciones de lingüística y fonética.

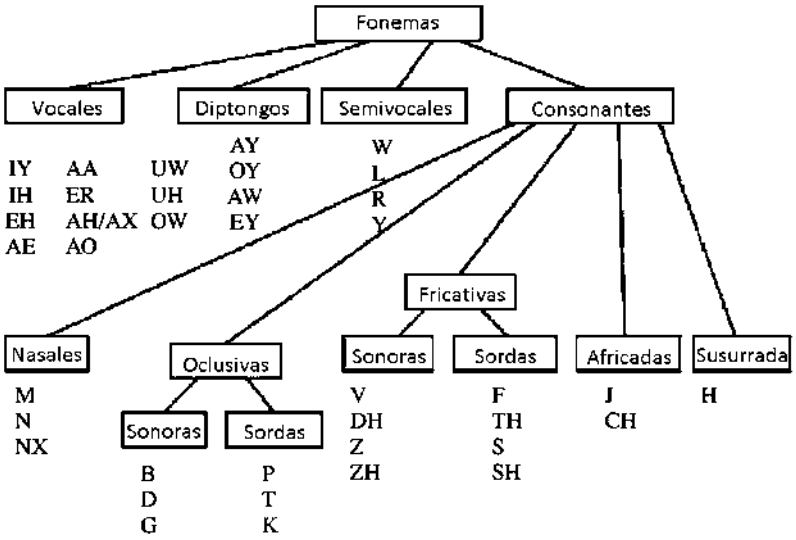

Figura 2.8: Fonemas en inglés americano. Fuente [Rabiner11].

En la Figura 2.8, los 39 fonemas que incluye son:

- 11 vocales
- 4 diptongos

- 4 semivocales
- 3 consonantes nasales
- 6 consonantes oclusivas sonoras y sordas
- 8 fricativas sonoras y sordas
- 2 consonantes africadas
- 1 sonido susurrado.

Los sonidos fricativos se producen por la fricción del aire al pasar entre dos órganos bucales, los sonidos oclusivos se producen por el cierre brusco momentáneo de alguna parte de la boca que impide la salida de aire, los sonidos de africados son el resultado de la articulación de la fricción y oclusión y los sonidos nasales se articulan dejando salir aire expirado por la nariz [Rabiner11].

Vocales

Si nos atenemos a las configuraciones del tracto y a la fuente de excitación que corresponden a cada fonema, una posible clasificación, los agrupa en vocálicos y consonánticos (se producen sin la vibración de las cuerdas vocales y por una excitación glótica ruidosa.). Esta división se sustenta tanto en las características acústicas como en los gestos articulatorios que dan lugar a cada tipo de sonido [Cobeta13].

En la articulación de sonidos vocálicos, el tracto muestra una configuración relativamente abierta y la fuente de excitación es siempre glótica. Las propiedades de estos sonidos persisten por un tiempo apreciable o cambian muy lentamente mientras se mantenga la configuración del tracto. Para este tipo de sonidos, los pulsos glóticos estimulan el tracto vocal que actúa como sistema resonador, este puede modificar su configuración y con ello sus frecuencias de resonancia (formantes), como si se tratara de un filtro acústico adaptativo. Esta posibilidad de variación es la que permite al parlante producir diferentes sonidos vocálicos. La forma del tracto en la producción de las vocales está controlada principalmente por la posición de la lengua, pero las posiciones de la mandíbula, los labios y, en menor medida, el velo también influye en el sonido resultante [Cobeta13, Rabiner11].

2.1.5. Teoría Fuente-Filtro

Existen diversas teorías para explicar la producción de la voz, las más importantes son la Teoría Fuente-Filtro, la Teoría Mioelástica-Aerodinámica y Mucoondulatoria [Cobeta13], y la Teoría Acústica: propagación del sonido en un tubo uniforme sin pérdidas (modelado a partir de ecuaciones diferenciales) [Rabiner11, Alzamendi16].

Las teorías Acústica, Mioelástica-Aerodinámica y Mucoondulatoria presentan la desventaja de ser matemáticamente intratables, excepto para los casos más simples y si se cuenta con información muy específica del aparato fonador, la cual muchas veces es difícil de obtener. Todo esto, hace que esta representación de la fonación resulte de poca utilidad para el análisis y procesamiento de señales de voz reales [Cobeta13, Alzamendi16].

La Teoría Fuente-Filtro (TFF) fue propuesta originalmente por Fant en la década de 1960 [Fant60], esta teoría, en su intento de simplificación, considera sólo tres elementos en la producción de la voz: la excitación (el flujo glótico modulado por la vibración de las cuerdas vocales), la transmisión (condicionada por la configuración y la resonancia del tracto vocal supraglótico) y la radiación (debida a la configuración de la apertura de la boca por la posición de los labios). El modelo asume el comportamiento lineal del sistema y la no variación en el tiempo del tracto vocal. Por lo tanto se puede considerar de forma independiente la excitación, la cavidad de resonancia y el efecto de radiación de los labios, permitiendo su análisis acústico y la extracción de los parámetros que posibilitan una aproximación al fenómeno fonatorio con suficiente fiabilidad [Cobeta13]. La Figura 2.9 ilustra este modelo simple.

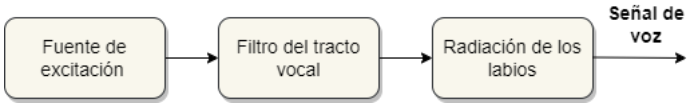

Figura 2.9: Diagrama a bloques de la Teoría Fuente-Filtro. Modificado de [deOliveira Dias12].

Por otro lado, la TFF al asumir una independencia entre sus componentes, desprecia toda interacción entre ellos, bajo la hipótesis de que no repercuten significativamente en el desempeño del modelo. Esta última hipótesis ha sido fuertemente criticada por la comunidad científica, a lo largo del tiempo. Diferentes autores han demostrado que la in-

fluencia entre el flujo de aire glótico y el tracto vocal es fuerte y extremadamente compleja [Chi07, Fant93, Titze08]. Sin embargo, la TFF se puede considerar suficiente para la mayoría de los casos de interés, lo que explica, por ejemplo, que se utilice ampliamente en los sistemas de procesamiento de voz [deOliveira Dias12].

Tomando en cuenta las ventajas y desventajas de cada una de las teorías que explican la producción de la voz, se selecciona para esta tesis la TFF, ya que esta estudia la fonación empleando conceptos propios del campo de señales y sistemas [Fant60]. Se caracteriza por presentar una formulación relativamente sencilla y por estar respaldada por un sólido marco conceptual. A su vez, ha demostrado ser fundamental en la evolución de las tecnologías involucradas en la comunicación, el procesamiento del habla y en la síntesis de voz [Alzamendi16].

Según la TFF, el sistema de producción de la señal de voz admite un modelado muy sencillo. En la Figura 2.10 se muestra una representación esquemática simplificada de la fonación de acuerdo con esta teoría. Se introduce un oscilador que genera un tren de impulsos de frecuencia controlada (equivalente a la frecuencia fundamental de la voz), junto con un generador de ruido blanco. Un conmutador permite seleccionar uno u otro tipo de señal, y con un sistema puede controlarse la ganancia o amplificación del proceso. Estos osciladores, junto con el conmutador, modelan el funcionamiento de la glotis en el ser humano. En este esquema, el tracto vocal se modela mediante un filtro resonante, cuya respuesta se controla a voluntad variando un conjunto de parámetros que gobiernan el comportamiento del filtro. La radiación de los labios/fosas nasales generalmente es denominada simplemente radiación de los labios, se modela con un filtro de paso alto aproximado por una derivada en el dominio del tiempo de primer orden [Cobeta13].

En el marco de la TFF, el filtro del tracto vocal, suele considerarse lineal e invariante en el tiempo, en ventanas o segmentos de señal de voz de corta duración ($20-40$ ms). Asimismo, se permite que este filtro varíe para cada ventana analizada. La razón de esto último es que los cambios en el tracto vocal son lentos, en comparación con la dinámica de la señal de voz [Rabiner11].

En el dominio del tiempo, la producción de la voz se puede representar por medio de una convolución de sus elementos (es decir, la fuente glótica, el filtro del tracto vocal

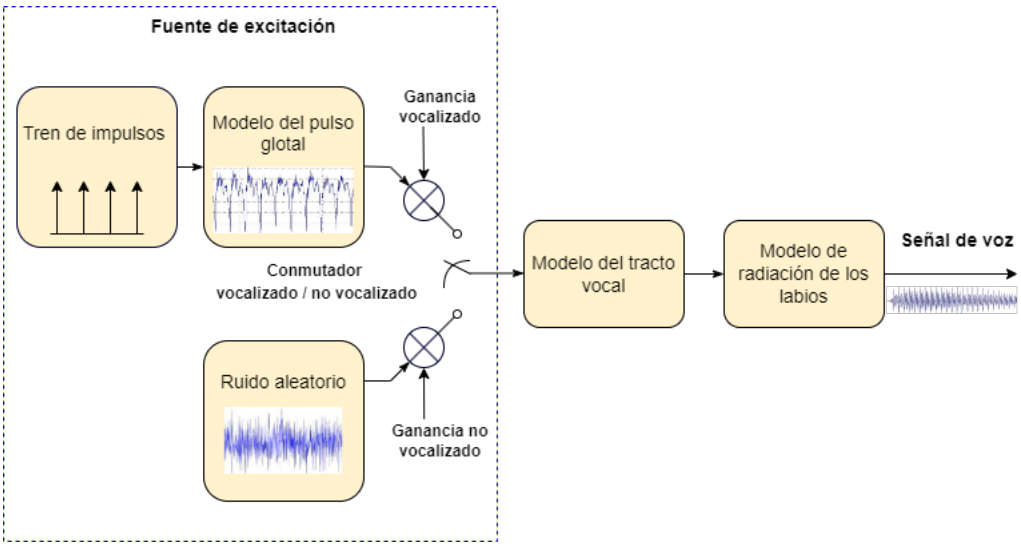

Figura 2.10: Modelo simplificado de la Teoría Fuente-Filtro. Modificado de [deOliveira Dias12].

y la radiación de los labios) [Degottex10]. Por lo tanto, matemáticamente la TFF para la producción de voz se puede expresar de la siguiente manera:

$$s[n] = u[n] * h[n] * r[n] \tag{2.1}$$

donde $s[n]$ es la señal de salida, es decir, la señal de voz, $u[n]$ es la señal de la fuente de excitación, $h[n]$ es la respuesta al impulso del tracto vocal y $r[n]$ es la radiación de los labios/fosas nasales.

En el dominio $\mathcal{Z}$, la Ecuación 2.1 se puede escribir como:

$$S(z) = U(z)H(z)R(z) \tag{2.2}$$

donde $U(z)$ es la transformada $\mathcal{Z}$ de la excitación acústica a nivel de la glotis.

Las resonancias y anti-resonancias del tracto vocal se combinan en un solo filtro $H(z)$, denominado filtro del tracto vocal y la radiación de los labios y las fosas nasales se combinan en un solo filtro $R(z)$, denominado radiación. Por lo tanto, el filtrado inverso de

la glotis requiere resolver la ecuación [deOliveira Dias12]:

$$U(z) = \frac{S(z)}{H(z)R(z)} \tag{2.3}$$

es decir, para determinar la forma de onda glótica, se debe eliminar la influencia del tracto vocal y la radiación de los labios/fosas nasales. En el caso de una señal de voz sonora, la forma de onda glótica presenta una forma periódica típica.

Por lo general, el filtro del tracto vocal se modela como un filtro de puros los polos de orden p:

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{2.4}$$

donde a_k son los coeficientes del filtro, los polos del filtro corresponden a resonancias del tracto vocal y, por tanto, a las frecuencias de los formantes del tracto vocal [Murphy08].

La radiación de los labios/fosas nasales impone un filtro de paso alto aproximado por una derivada en el dominio del tiempo de primer orden [Murphy08], lo que significa que la derivada del flujo glótico es la excitación efectiva del tracto vocal. Por lo tanto:

$$R(z) = 1 - dz^{-1} \tag{2.5}$$

donde d es el coeficiente de radiación labios/fosas nasales, cuyo valor es cercano a 1 [Javkin87]. Por lo general, d es un valor entre 0.95 y 0.99 para que el cero se encuentre dentro del círculo unitario en el plano $\mathcal{Z}$.

Esta ecuación se puede escribir como [Javkin87]:

$$R(z) = \frac{1}{\sum_{k=0}^{N} d^k z^{-k}} \tag{2.6}$$

donde N es teóricamente infinito, pero en la práctica finito porque $d < 1$. Este resultado sugiere que el efecto de un cero puede aproximarse mediante un número suficientemente grande de polos.

2.2. Análisis de Predicción Lineal

La señal de voz es producto de los cambios de presión que se transmiten a través del aire. Su evolución temporal no es imprevisible a corto plazo; los instantes pasados dan una idea aproximada del futuro, debido a que su evolución no es abrupta sino suave (aunque sea más o menos rápida). Hay, pues, una dependencia entre pasado y futuro. Ahora bien, toda información predecible es redundante. Según el teorema de la información de Shannon [Shannon49], cuanto más predecible sea un suceso menor cantidad de información aporta. Por otra parte, podemos realizar aproximaciones al patrón original sin que por ello se pierda la información subyacente que nos permita realizar la evaluación. ¿Para qué caracterizar algo con una cantidad n de datos si para el problema en estudio podemos mantener rasgos identificativos con menos volumen de información? Por lo tanto, la finalidad de la parametrización debe ser eliminar toda redundancia informativa, manteniendo las características y rasgos de la señal original que permitan una evaluación con el mínimo número de parámetros [Cobeta13].

El análisis de predicción lineal, nos permite calcular la envolvente espectral y los coeficientes LPC (Linear Prediction Coefficients, por sus siglas en inglés). Se utiliza para la extracción de rasgos y es la base de la mayoría de los sistemas de codificación de voz. Su popularidad se debe en gran parte a su sencilla formulación y baja demanda de cálculos en comparación con otras técnicas. La base matemática de este método ha sido profusamente investigada y se ha utilizado en gran número de aplicaciones dentro del procesamiento de voz, pues permite estimar, de manera precisa y relativamente rápida, parámetros como el espectro, los formantes, la frecuencia fundamental o la morfología del pulso glótico [Cobeta13].

La codificación lineal predictiva está basada en la idea de que una muestra de voz puede predecirse mediante una combinación lineal de las muestras anteriores de voz dentro de un cierto intervalo, para poder hacer eso, es necesario determinar los coeficientes predictores, esto se logra minimizando la suma de los cuadrados de las diferencias entre las muestras obtenidas mediante la predicción y las muestras reales [Ibarrola11].

Como vimos en la Figura 2.10, la TFF obtiene la señal de voz como resultado

de introducir una señal de excitación (diferente según se trate de un segmento vocalizado o no vocalizado) a la entrada del filtro que modela el tracto vocal y la radiación de los labios/fosas nasales.

El tracto vocal es modelado por un filtro de puros polos [Rabiner11], cuya función de transferencia se presenta en la Ecuación 2.4, si le aplicamos transformada $\mathcal{Z}$ inversa, se obtiene una simple ecuación de diferencias de orden p:

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + Gu[n] \tag{2.7}$$

donde a_k son los coeficientes del filtro del tracto vocal, $u[n]$ es la señal de excitación y G es la ganancia de la excitación.

Como se dijo antes, un predictor lineal predice una muestra de voz en base a las p muestras anteriores, así:

$$\tilde{s}[n] = \sum_{k=1}^{p} \alpha_k s[n-k] \tag{2.8}$$

donde α_k son los coeficientes del predictor.

Según Rabiner [Rabiner11], la función del sistema de este predictor lineal de orden p es el polinomio de transformación $P(z)$

$$P(z) = \sum_{k=1}^{p} \alpha_k z^{-k} = \frac{\tilde{S}(z)}{S(z)}. \tag{2.9}$$

$P(z)$ a menudo se denomina polinomio predictor. Al hacer esta aproximación se produce un error de predicción $e(n)$, este es la diferencia entre las muestras reales de voz y las muestras que produce el predictor:

$$e[n] = s[n] - \tilde{s}[n] = s[n] - \sum_{k=1}^{p} \alpha_k s[n-k]. \tag{2.10}$$

Al comparar (2.7) y (2.8) concluimos que para poder afirmar que los coeficientes predictores y los parámetros del tracto vocal coinciden, es decir $\alpha_k = a_k$ entonces debería cumplirse $e[n] = Gu[n]$ lo cual significa que $e[n]$ es una buena aproximación de la fuente de

excitación de un sistema de producción de voz. El error de predicción cuadrático total de tiempo corto se define como [Rabiner11]:

$$E_n = \sum_m e_n^2[m] = \sum_m \left(s_n[m] - \sum_{k=1}^{p} \alpha_k s_n[m-k] \right)^2, \tag{2.11}$$

donde $s_n[m]$ es la m-ésima muestra contada desde el inicio del marco que comienza en la n-ésima muestra, es decir la muestra $s[n+m]$.

Para encontrar los valores de α_k que minimizan E_n, hacemos $\partial E_n/\partial \alpha_i = 0$ para todo $i = 1, 2, ..., p$ [Ibarrola11]. Resolviendo y simplificando tenemos que:

$$\sum_{k=1}^{p} \alpha_k \phi_n[i,k] = \phi_n[i,0], \qquad i = 1, 2, ..., p \tag{2.12}$$

donde

$$\phi_n[i,k] = \sum_m s_n[m-i] s_n[m-k]. \tag{2.13}$$

Para encontrar los coeficientes predictores que minimizan el error se deben calcular los coeficientes $\phi_n[i,k]$ para todo $1 \leq i \leq p$ y $0 \leq k \leq p$ y luego resolver el sistema de ecuaciones dado por (2.12). Para lograr esto, existen varios métodos, entre ellos el método de la covarianza, y el de la autocorrelación, describiremos aquí este último. En las ecuaciones anteriores, los límites de las sumatorias que utilizan a m como índice se han dejado sin especificar, sin embargo, al referirnos al error de predicción de tiempo corto, el intervalo coincide con el ancho N de los marcos, así, $s[n+m]w[m]$ es cero fuera del intervalo $[0, N-1]$ y el error de predicción de tiempo corto solo tiene valores distintos de cero dentro del intervalo $[0, N+p-1]$, por esta razón (2.12) se convierte en [Rabiner11]:

$$\phi_n[i,k] = \sum_{m=0}^{N+p-1} s_n[m-i] s_n[m-k], \quad i = 1, 2, ..., p, \ k = 0, 1, ..., p. \tag{2.14}$$

Haciendo $r = m - i$, tenemos que $m - k = r + i - k$

$$\phi_n[i,k] = \sum_{r=0}^{N-1-(i-k)} s_n[r]s_n[r+i-k], \quad i = 1,2,...,p, \ k = 0,1,...,p. \tag{2.15}$$

En la ecuación anterior (2.15), podemos apreciar que $\phi_n(i,k)$ es idéntica a la función de autocorrelación evaluada en $(i-k)$. En vista de que la función de autocorrelación es una función par, se cumple que [Ibarrola11]:

$$\phi_n[i,k] = R_n[|i-k|] \tag{2.16}$$

Entonces (2.12) se puede expresar como [Ibarrola11]:

$$\sum_{k=1}^{p} \alpha_k R_n[|i-k|] = R_n[i], \qquad i = 1,2,...,p. \tag{2.17}$$

2.2.1. Solución de las ecuaciones LPC

Para encontrar los coeficientes LPC, debemos resolver el sistema de ecuaciones expresado en forma compacta mediante (2.17), este sistema de ecuaciones podemos expresarlo como $T\alpha = r$ o en forma desarrollada [Ibarrola11]:

$$\begin{bmatrix} R_n[0] & R_n[1] & R_n[2] & ... & R_n[p-1] \\ R_n[1] & R_n[0] & R_n[1] & ... & R_n[p-2] \\ R_n[2] & R_n[1] & R_n[0] & ... & R_n[p-3] \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ R_n[p-1] & R_n[p-2] & R_n[p-3] & ... & R_n[0] \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ . \\ . \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R_n[1] \\ R_n[2] \\ R_n[3] \\ . \\ . \\ R_n[p] \end{bmatrix} \tag{2.18}$$

La matriz de autocorrelación T es una matriz de Topeplitz dado que es simétrica y los valores a lo largo de cualquier diagonal son un mismo valor. Los métodos mas conocidos para solucionar un sistemas de ecuaciones con estas características son los de Levinson y Robinson, también conocido como procedimiento recursivo de Durbin [Rabiner11, Ibarrola11].

Método recursivo de Durbin para solucionar las ecuaciones de autocorrelación

Este procedimiento comienza con un predictor de primer orden, es decir, de un solo coeficiente e incrementa el orden recursívamente utilizando las soluciones de orden inferior para obtener soluciones de órdenes superiores y así sucesivamente hasta llegar al orden deseado. La notación α_i^p denota al i-ésimo coeficiente de un predictor de orden p [Ibarrola11]. El algoritmo de Levinson-Durbin se detalla en el Algoritmo 1.

Algoritmo 1 Algoritmo de Levinson-Durbin. Fuente [Rabiner11]

Entrada: R_n, p
Salida: α
$E^{(0)} = R_n(0)$;
para $i \leftarrow 1$ **a** p **hacer**
  $k_i = \left[R_n(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R_n(i-j)\right] / E^{(i-1)}$;
  $\alpha_i^{(i)} = k_i$;
  si $i > 1$ **entonces**
    para $j \leftarrow 1$ **a** $i-1$ **hacer**
      $\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)}$;
    fin
  fin
  $E^{(i)} = (1 - k_i^2) E^{(i-1)}$;
fin
$\alpha_j = \alpha_j^{(p)}, \qquad j = 1, 2, ..., p$
devolver α

En el proceso de calcular los coeficientes predictores para un predictor de orden p se deben calcular los coeficientes de todos los predictores de órdenes inferiores a p. El error de predicción se puede estar monitoreando así como los coeficientes PARCOR (k_i) [Rabiner11, Ibarrola11].

2.3. Preprocesamiento de la señal de voz

La señal de voz por lo general se encuentra almacenada en archivos de audio. Estos audios no solo contienen la señal de audio deseada, también contienen tiempos muertos (sin señal de audio presente) al inicio y al final de la elocución. Para solucionar este problema existen técnicas como: régimen de cruces por cero de tiempo corto y energía de tiempo

corto, las cuales permiten detectar el inicio y final de solo la señal de audio (discriminar entre silencio y voz).

El pulso glotal siempre está presente en los sonidos vocalizados. Las propiedades de estos sonidos persisten por un tiempo apreciable o cambian muy lentamente mientras se mantenga la configuración del tracto vocal. Como se menciono antes, en el marco de la TFF, el filtro del tracto vocal, $H(z)$, suele considerarse lineal e invariante en el tiempo, en segmentos (marcos) de señal de voz de corta duración ($20 - 40$ ms). Es por esto qué es importante dividir en marcos a la señal de voz e identificar cuales de estos contienen sonidos vocalizados.

2.3.1. Régimen de cruces por cero de tiempo corto

Los cruces por cero ocurren cuando muestras sucesivas de la señal de audio tienen signo distinto. El régimen de cruces por cero de tiempo corto es una medida de frecuencia en una señal. Los fragmentos de audio que contienen voz normalmente tiene un régimen inferior que las fragmentos del audio que no contienen voz (excepto en los audios con supresión de ruido, ya que sucede el efecto contrario). Por lo tanto, el régimen de cruces por cero puede ser utilizado para diferenciarlas. El proceso consiste en tomar pequeños fragmentos del audio y calcular su régimen de cruces por cero, que es comparado con un umbral para identificar si el fragmento contiene voz. En la siguiente ecuación se muestra como se calcula el régimen de cruces por cero de un fragmento del audio de N muestras [Rabiner11].

$$Zn = \frac{1}{2N} \sum_{i=1}^{N} |\text{signo}(s[i]) - \text{signo}(s[i-1])|, \tag{2.19}$$

donde a Zn se le conoce como régimen de cruces por cero de tiempo corto.

La función signo recibe una muestra y si es mayor o igual que 0 regresa 1, en caso contrario regresa -1. Cuando las muestras tienen el mismo signo el resultado es cero, pero en caso contrario el resultado es dos, por este motivo el resultado del sumatorio es divido por dos [Rabiner11]. La función signo se describe por la Ecuación 2.20.

$$\text{signo}(s[i]) = \begin{cases} 1 & , \quad s[i] \geq 0 \\ -1 & , \quad s[i] < 0 \end{cases} \tag{2.20}$$

2.3.2. Energía de tiempo corto

La energía de tiempo corto es la energía de una ventana o segmento de señal de voz. Es la sumatoria del cuadrado de cada muestra de una ventana de tamaño N. La energía de tiempo corto es de ayuda en este caso para diferenciar los fragmentos que contienen voz, ya que los fragmentos que contienen voz tienen mayor energía que los fragmentos que no la contienen, como el ruido de fondo (no se ve afectada por la supresión de ruido) [Rabiner11]. En la Ecuación 2.21 se muestra como se calcula la energía de tiempo corto.

$$En = \sum_{i=1}^{N} s^2[i], \tag{2.21}$$

donde a En se le conoce como energía de tiempo corto.

2.3.3. División de audio en segmentos

La señal de voz se divide en segmentos (marcos) de 30 ms, para esto se utilizan ventanas. Esta ventana es desplazada por toda la señal de audio. El desplazamiento es de un tercio del tamaño de la ventana (10 ms) [Rabiner11]. En la Figura 2.11 se muestra la configuración de la división del audio en ventanas.

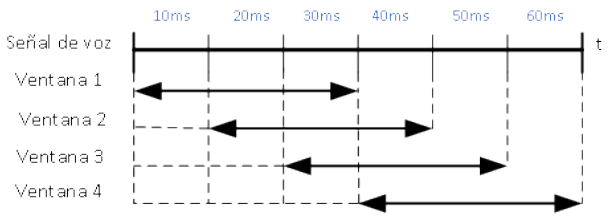

Figura 2.11: Configuración de la división del audio en ventanas. La señal de voz es dividida utilizando una ventana que toma fragmentos de 30 ms y es desplazada por toda la señal cada 10 ms.

2.3.4. Identificación de sonidos vocalizados usando los coeficientes LPC

La señal de error $e[n]$ obtenida en el análisis LPC puede ser usada para estimar el tono directamente de la señal de voz y esta definida como:

$$e[n] = s[n] - \tilde{s}[n] = s[n] - \sum_{k=1}^{p} \alpha_k s[n-k] = Gu[n] \tag{2.22}$$

La señal de error es una buena aproximación de la fuente de excitación de un sistema de producción de voz, el cual es modelado por un predictor de orden p. Basado en este razonamiento, es de esperarse que el error de predicción aumente drásticamente cada vez que inicia otro periodo del tono para sonidos vocalizados, como podemos apreciar en la Figura 2.12. Por lo tanto, dicho periodo se puede determinar detectando las posiciones de las muestras de la señal $e[n]$ que son de mayor amplitud y midiendo entonces la diferencia de tiempo que hay entre muestras de $e[n]$ que superen un valor umbral. Alternativamente, se puede estimar el tono determinando la autocorrelación de la señal de error y detectando la posición del pico mas alto y si supera un valor umbral que es determinado empíricamente (en esta tesis es de 0.2), es un sonido vocalizado. En esta tesis para obtener el tono se utiliza la autocorrelación, ya que amplifica las muestras de la señal que son periódicas y las que no las atenúa, facilitando la detección de los sonidos vocalizados [Rabiner11].

La razón por la que la señal de error es tan útil para la estimación del tono es porque el espectro de la señal de error es prácticamente plano ya que los formantes han sido eliminados y no aparecen en la señal de error [Rabiner11].

2.4. Estimación del Pulso Glotal

La importancia del análisis del flujo glótico está muy bien establecida en diferentes áreas pero antes se planteó que los procedimientos médicos que permiten la observación de la vibración de las cuerdas vocales son incómodos para el parlante e interfieren con el comportamiento normal de fonación. La estimación del flujo glótico a partir de una señal de voz ha sido un desafío en las últimas décadas y, como consecuencia, se han propuesto varias técnicas para la estimación del pulso glotal y varios modelos glóticos para definir

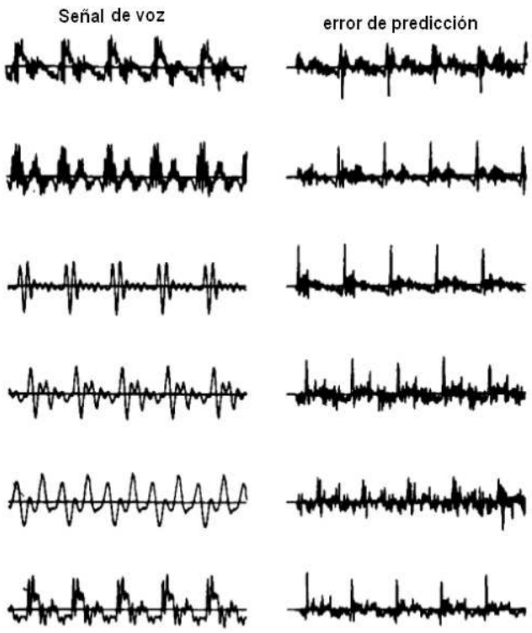

Figura 2.12: Ejemplos de la señal de error de predección. Donde se puede apreciar que la señal de error es prácticamente plana en comparación con la señal de voz. Fuente [Ibarrola11].

analíticamente un período del flujo glótico.

Los modelos glóticos para definir analíticamente un período del flujo glótico más conocidos y utilizados son: Rosenberg [Rosenberg71], Fant [Fant79], Liljencrants-Fant (modelo LF) [Fant85] y Transformada-LF (LF^{Rd}) [Fant95]. Estos modelos utilizan principalmente un conjunto de instantes de tiempo, como se muestra en la Figura 2.13:

t_c: duración del periodo ($t_c = T_0 = 1/f_0$)

t_p: tiempo del máximo del pulso;

t_e: tiempo del mínimo de la derivada del pulso;

t_a: duración de la fase de retorno.

2.4.1. Métodos para estimar el Pulso Glotal

En las últimas décadas, se han desarrollado varios métodos y técnicas para la estimación de la forma de onda glótica durante el habla sonora y continúa siendo un campo

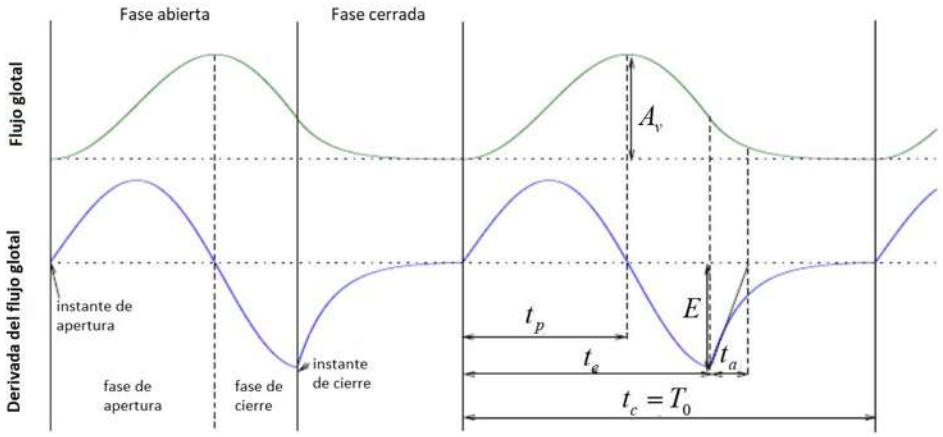

Figura 2.13: Esquema principal del pulso glotal utilizado por la mayoría de los modelos glóticos. Fuente [deOliveira Dias12].

de investigación activo en la actualidad. Algunos métodos trabajan en el dominio del tiempo, otros en el dominio de la frecuencia y algunos otros utilizan modelos matemáticos del pulso glotal, como los mencionados anteriormente.

A continuación se dará una breve explicación de los métodos mas utilizados en la literatura para la estimación del pulso glotal:

- Filtrado Inverso Adaptativo Iterativo (IAIF, por sus siglas en inglés), es un método de filtrado inverso semiautomático propuesto por Alku [Alku92b]. El método utiliza una señal de la voz como entrada y genera una estimación de la señal de flujo glotal correspondiente. Este procedimiento tiene tres partes fundamentales: análisis, filtrado inverso e integración. La contribución glótica al espectro del habla se estima inicialmente utilizando una estructura iterativa. Esta contribución se cancela y, luego, se modela la función de transferencia del tracto vocal. Finalmente, la excitación glótica se estima cancelando los efectos del tracto vocal (usando filtrado inverso) y la radiación labial (por integración). El método opera en dos iteraciones: la primera fase, realiza una estimación de la función del tracto vocal y aplica filtrado inverso a la señal con esa estimación, y genera una estimación de la fuente glótica que se utiliza como entrada de la segunda fase para lograr una estimación más precisa.

- Filtrado Inverso Adaptativo Iterativo Síncrono del Tono (PSIAIF, por sus siglas en

inglés) [Alku92b, deOliveira Dias12], es un método de filtrado inverso basado en IAIF. El pulso glótico se obtiene aplicando el método IAIF dos veces a la señal de voz. El resultado del primer procedimiento IAIF se utiliza para calcular el período fundamental que es importante para calcular la nueva ventana, antes de volver a aplicar IAIF.

- Filtrado Inverso Iterativo (Modelo de Murphy) [Murphy08], es una versión iterativa del método de filtrado inverso ideado por Alku [Alku92a]. Primero, se elimina el efecto de la radiación de los labios de la señal de voz. Segundo, se obtiene una primera estimación de la función inversa del pulso glótico simple, que se utiliza para cancelar el comportamiento del pulso glotal sobre la voz compensada por radiación, produciendo una señal de voz desglotalizada. Tercero, se deriva el modelo para el tracto vocal mediante el filtrado inverso de la voz desglotalizada con filtros de red (lattice filters) y extrayendo el modelo del tracto vocal. Cuarto, toma el modelo inverso del tracto vocal, y lo aplica a la voz compensada por radiación, que se extrajo en el primer paso y genera una señal residual que contiene información sobre la segunda derivada del pulso glótico. Quinto, se deriva el modelo de pulso glotal y se repite los pasos del segundo al quinto en una iteración en bucle hasta un máximo de iteraciones o hasta que la señal de la derivada del pulso glotal cambie demasiado poco.

- Ceros de la Transformada $\mathcal{Z}$ (ZZT, por sus siglas en inglés) [Bozkurt05]. El algoritmo toma la transformada $\mathcal{Z}$ de un segmento de voz sincronizada con el instante de cierre glótico (GCI, por sus siglas en inglés) y se calculan las raíces (ceros). Se identifican tres conjuntos de ceros. El conjunto de ceros que tienen un módulo igual a uno representa el tren de impulsos subyacente a la periodicidad de la voz (es decir el tono). El conjunto de ceros que tienen un módulo mayor que uno se debe a la parte anticausal de la fuente de voz. El conjunto de ceros que tienen módulo menor que uno, se debe a la parte causal de la fuente de voz. Por lo tanto, el uso de la transformada de Fourier discreta (DFT, por sus siglas en inglés) y de la transformada inversa de Fourier discreta (IDFT, por sus siglas en inglés) para cada uno de estos grupos permite estimar la fuente glótica y el filtro del tracto vocal.

- En 2009, Drugman [Drugman09] propuso un método, llamado “Descomposición Com-

pleja Basada en Cepstrum"(CC, por sus siglas en inglés), basado en los mismos principios de la descomposición ZZT, es decir, la señal de voz es una señal de fase mixta donde la contribución de fase máxima está relacionada con la fase abierta glótica. y la fase mínima está relacionada con el cierre de la glotis y el componente del tracto vocal. Este enfoque tiene una clara ventaja: computacionalmente es mucho más rápido que el ZZT.

- Método de JAVKIN [Javkin87]. Se desarrolló en el dominio de la frecuencia porque, según los autores, los formantes introducidos por el tracto vocal, así como el efecto de la radiación de los labios, se entienden mejor en este dominio que en el dominio del tiempo. Dado que el flujo glótico estimado es una función del tiempo, la Transformada $\mathcal{Z}$ se usa para convertir entre estos dos dominios. Debido a las interacciones entre el tracto vocal y la fuente, las frecuencias de los formantes y los anchos de banda se modulan durante la fase abierta del ciclo glótico. Luego, se debe obtener una estimación confiable de los parámetros del tracto vocal durante la fase de cierre glótico, que puede ser detectado a partir de la señal residual de LPC. Para eliminar los formantes y los efectos de la radiación del labio, es necesario construir un filtro que tenga la respuesta inversa. Se propone un filtro digital que modela el tracto vocal y se obtiene un modelo para cada formante.

- Filtrado inverso con criterio de variación mínima [Veeneman85]. Utiliza la señal de voz y un criterio de error normalizado mínimo para la determinación de los límites de la fase cerrada. Se determina el filtro que modela el tracto vocal para hallar la excitación del sistema se aplica el filtrado inverso sobre el tramo de señal. Finalmente, se elimina el efecto de radiación integrando la señal para obtener el pulso.

- Estimación del Pulso Glotal en el Dominio de la Frecuencia (FD-GPE, por sus siglas en inglés) [Dias12], este método trabaja en el dominio de la frecuencia, el cual analiza una región de un sonido sonoro para extraer información armónica y un modelo de envolvente espectral suavizada. La información armónica incluye las frecuencias, magnitudes y fases de todos los componentes sinusoidales que son armónicos de la frecuencia fundamental. Permite la manipulación independiente de la señal de mag-

nitud y fase, lo que agrega flexibilidad en el procesamiento de las contribuciones de magnitud y fase de ambas fuentes glóticas y filtro de tracto vocal.

2.4.2. Método propio para la estimación del Pulso Glotal

Nuestro método para extraer el flujo glotal (pulso glotal) es iterativo, similar a las técnicas de Murphy [Murphy08] o IAIF de Alku [Alku92a]. Estos métodos se basan en la TFF de Fant [Fant60] y se conocen como métodos de filtrado inverso.

La TFF establece que la excitación glótica, el filtro del tracto vocal y la radiación de los labios/fosas son independientes y, por lo tanto, linealmente separables de la señal de voz, si se conoce la función de transferencia del filtro del tracto vocal, se puede construir un filtro inverso para estimar la fuente de voz. En este caso el filtrado inverso trata de estimar la excitación glótica cancelando los efectos espectrales del tracto vocal y la radiación de los labios/fosas nasales. Si la producción de voz se describe como una convolución, el filtrado inverso puede entenderse como una operación de deconvolución [deOliveira Dias12]. En la Figura 2.14 se muestra el modelo de producción de voz de acuerdo a la TFF y el proceso de filtrado inverso para producir una estimación de la forma de onda del flujo glotal.

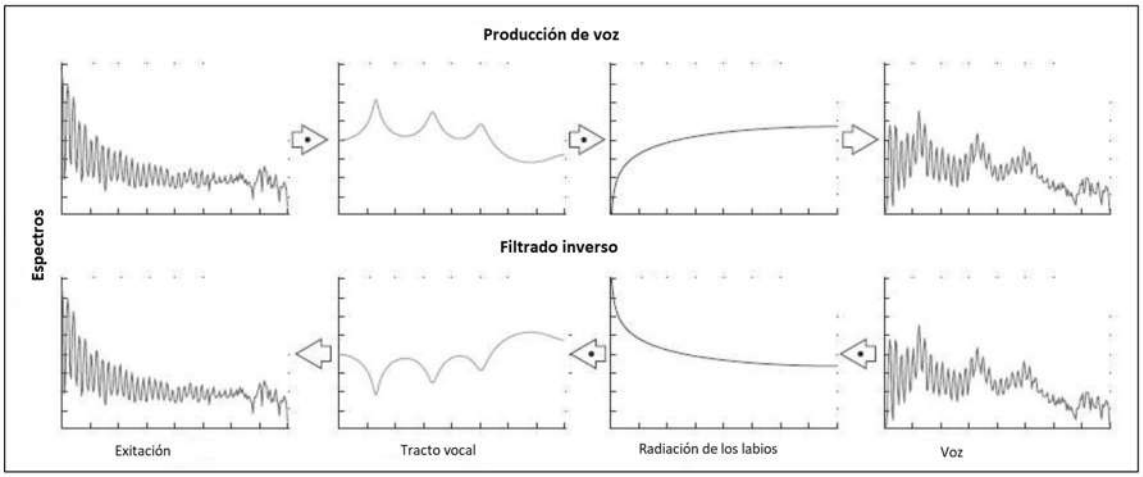

Figura 2.14: Espectros esquemáticos de la producción de voz. La parte superior de la imagen representa el modelo de producción de voz. La parte inferior de la imagen representa el proceso de filtrado inverso correspondiente. Fuente [deOliveira Dias12].

El método propuesto utiliza una señal de voz como entrada y genera una estimación

de la señal de flujo glotal correspondiente. Se calculan de manera iterativa los coeficientes que modelan el tracto vocal para lograr una estimación más precisa, para luego eliminar de la voz de entrada los efectos del tracto vocal por medio del filtrado inverso y de esta manera obtener la estimación del pulso glotal. En nuestro método no se elimina la radiación de los labios de la señal de voz, dado que esta acentúa más la información de las frecuencias altas, permitiendo un modelo más preciso en la pendiente del espectro del pulso glotal y conservando una mayor información de los órganos responsables de la fonación. Además de no utilizar un modelo matemático para caracterizar al pulso glotal (como si lo hace Murphy [Murphy08]), lo que reduce el costo computacional y se evita tener que seleccionar y analizar alguno de los modelos que caracterizan al pulso glotal.

Nuestra técnica de estimación del flujo glotal, es representado en la Figura 2.15, donde las líneas continuas representan el flujo de control y las líneas punteadas representan el flujo de datos (es decir, señales). A continuación se proporciona una descripción de cada paso:

- Se encuentran los parámetros $\alpha_1, ..., \alpha_p$ del filtro que modela al tracto vocal, estos se calculan por medio del análisis LPC a la señal de entrada $s^{(i)}[n]$, ya que en este punto $i = 0$, se debe considerar como $s^{(0)}[n]$.

- Al aplicar a $s^{(i)}[n]$ un filtro que es el inverso de este filtro que modela el tracto vocal, obtenemos una primera estimación de la onda de sonido al comienzo del tracto vocal, que es la forma de onda de flujo glotal denotada como $u^{(i)}[n]$. Hay que tener en cuenta que el primer análisis LPC obtuvo parámetros $\alpha_1, ..., \alpha_p$ que no dependen exclusivamente de la forma del tracto vocal sino también de la forma de onda del flujo glótico.

- Para determinar parámetros que dependen únicamente del tracto vocal necesitamos desglotalizar la voz, lo hacemos simplemente restando la estimación del flujo glótico $u^{(i)}[n]$ de la voz $s^{(i)}[n]$ de la cual se obtuvieron los parámetros del filtro que modela el tracto vocal. De esta forma obtenemos la llamada "voz desglotalizada", luego re-emplazamos la señal de voz $s^{(i)}[n]$ por $s^{(i+1)}[n]$ y repetimos el análisis LPC; filtrado

inverso; y desglotalización, hasta un máximo de iteraciones o hasta que la señal de voz desglotalizada cambie demasiado poco.

- Para medir el cambio en la señal de voz desglotalizada, se calcula la energía de la diferencia de dos formas de onda de voz desglotalizadas consecutivas.

- Una vez que se completa este proceso, se cuenta con los parámetros $\alpha_1, ..., \alpha_p$ que dependen exclusivamente de la forma del tracto vocal. Se aplica a la señal de entrada $s^{(0)}[n]$ un filtrado inverso con los parámetros que dependen exclusivamente de la forma del tracto vocal, obteniendo la estimación final del flujo glotal.

2.5. Conclusiones del capítulo

El pulso glotal se genera gracias a la interacción de los distintos músculos, ligamentos, cartílagos, membranas y estructuras de la laringe (la longitud de las cuerdas vocales, el perfil dinámico tridimensional de la glotis, etc.), es por esto que el pulso glotal cuenta con características únicas para cada parlante y puede ser usado para la identificación texto-independiente de parlantes. Para la estimación del pulso glotal hay que encontrar los fragmentos de voz con sonido vocalizado, ya que las vocales siempre tienen como fuente de excitación el pulso glotal. Se implementa un método propio para la estimación del pulso glotal, el cual cuenta con las siguientes ventajas sobre otros métodos: no necesita de una detección del GCI y sincronización con el tono, esto es conveniente ya que para algunas voces es difícil encontrar el GCI y en ocasiones no se puede encontrar; aunado a esto, presenta la ventaja de que sus trenes de pulsos glotales tienen una longitud igual (se utilizan segmentos de 30 ms), esto es adecuado para la utilización de redes neuronales artificiales. Además de lo mencionado anteriormente, es muy robusto, ya que itera hasta encontrar una buena estimación del tracto vocal y del pulso glotal, con un menor costo computacional comparado con otras técnicas.

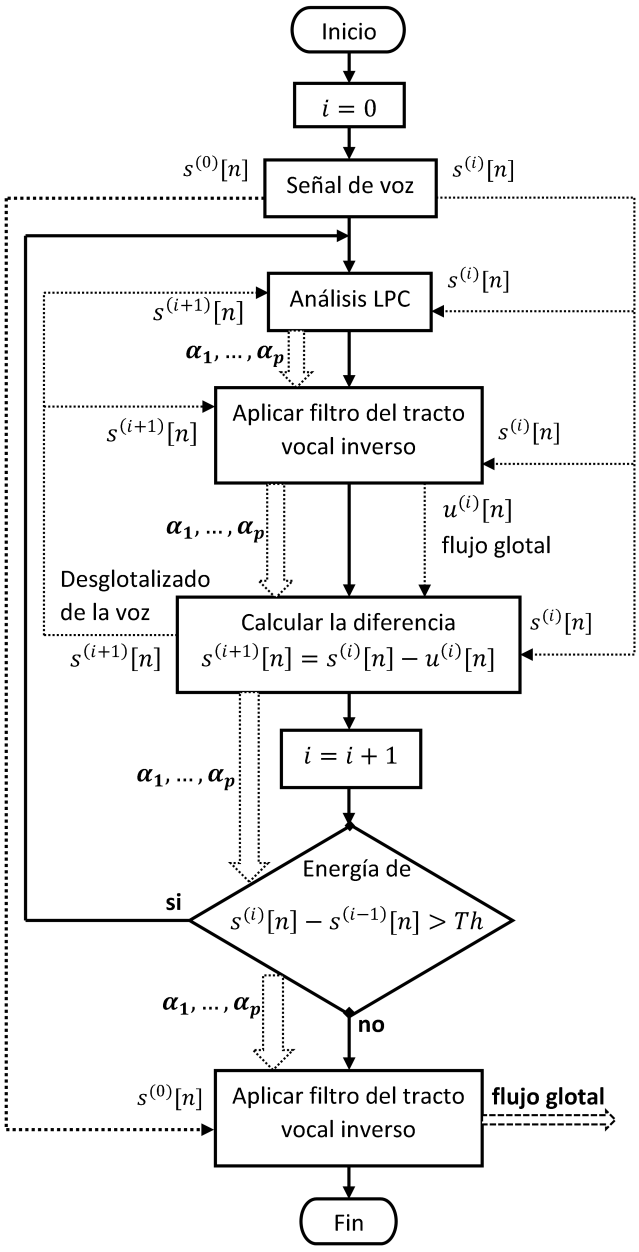

Figura 2.15: Nuestra técnica de estimación del pulso glotal. Las líneas continuas representan el flujo de control y las líneas punteadas representan el flujo de datos (es decir, señales).

Capítulo 3

Redes Neuronales Convolucionales para la Identificación de Parlantes

Las Redes Neuronales Convolucionales 2D (CNN 2D, por sus siglas en inglés) han tenido un gran éxito en problemas de visión artificial, como en la clasificación y segmentación de imágenes, entre otras aplicaciones. Sin embargo, las CNN también pueden ser aplicadas para la clasificación de series de tiempo o señales de audio en 1D, así como para la clasificación de datos volumétricos usando convoluciones en 3D.

Las CNN, son capaces de aprender a clasificar todo tipo de datos distribuidos de una forma continua a lo largo del mapa de entrada, y a su vez sean estadísticamente similares en cualquier lugar del mapa de entrada. Por esta razón, son especialmente eficaces para clasificar imágenes, por ejemplo para el auto-etiquetado de imágenes.

Para poder entender como funcionan las CNN es necesario conocer algunos conceptos básicos sobre redes neuronales.

3.1. Nociones sobre Redes Neuronales

Es importante entender la diferencia entre Inteligencia Artificial (IA), Aprendizaje Automático (Machine Learning) y Aprendizaje Profundo (Deep Learning). En general el Aprendizaje Profundo es una subparte de una de las áreas de la IA conocida como Apren-

dizaje Automático [Torres18].

3.1.1. Redes Neuronales Artificiales

Las Redes Neuronales Artificiales (ANN, por sus siglas en inglés) están basadas en el funcionamiento de las redes de neuronas biológicas. Un cerebro humano está compuesto por una enorme cantidad de neuronas que interactúan entre ellas a través de una intrincada red de conexiones. En general una neurona consta de tres partes principales: las dendritas, el núcleo y el axón. Las dendritas se encargan de captar los impulsos nerviosos que emiten otras neuronas. Estos impulsos, se procesan en el núcleo y se transmiten a través del axón que emite un impulso nervioso hacia las neuronas contiguas. La sinapsis es un espacio que está ocupado por unas sustancias químicas denominadas neurotransmisores y está localizado entre las zonas del axón y las dendritas de las neuronas siguientes. Estos neurotransmisores son los que se encargan de bloquear o dejar pasar las señales que provienen de las otras neuronas [Caicedo09].

La neurona artificial está compuesta por las entradas (X), por los pesos sinápticos (W) que ponderan a las entradas (X), así como el sesgo (b) o umbral de una neurona y la función de activación (f_a). La neurona multiplica las entradas X por los pesos W para posteriormente agregar el sesgo b. El resultado de esta suma ponderada es tomado como entrada para la función de activación que regresa Y. En la Figura 3.1 se puede apreciar una analogía entre las neuronas biológicas y las artificiales, donde los pesos de la neurona artificial corresponden a la fuerza de las sinapsis, mientras que la suma, el sesgo y la función de activación corresponden al cuerpo de una neurona biológica, por su parte la salida y corresponde a la señal que pasa por el axón [Hagan14].

La arquitectura más simple que puede tener una red neuronal, se le conoce como Perceptrón y matemáticamente una ANN está dada por la siguiente ecuación:

$$Y_j = f_a(W^T X_j + b_j).$$

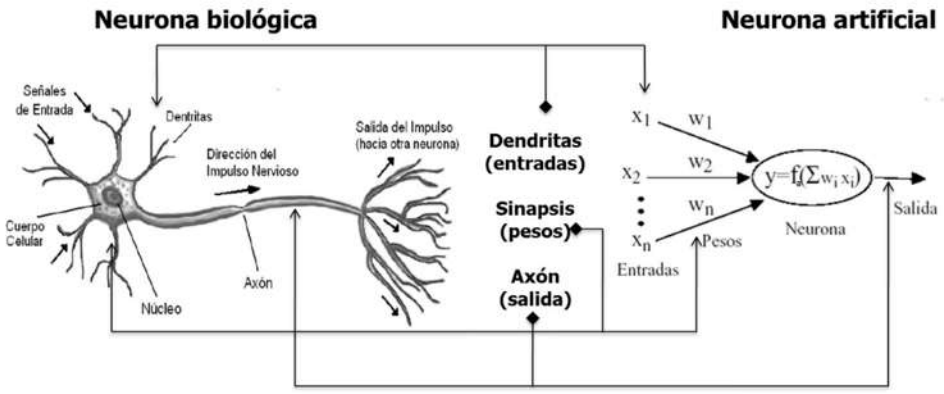

Figura 3.1: Comparación neurona artificial y neurona biológica. Fuente [Lao León17].

3.1.2. Redes Neuronales Multicapa

El Perceptrón es la versión más simple de red neuronal porque consta de una sola capa que contiene una sola neurona. Sin embargo, solo se utilizan para problemas muy sencillos. Para problemas más complejos se utilizan redes neuronales compuestas de numerosas capas y que cada una de ellas contenga muchas neuronas que se comunican con las de la capa anterior para recibir información, y estas a su vez comunican su información a las neuronas de la capa siguiente [Torres18].

Se le conoce como Perceptrón Multicapa (MLP, por sus siglas en inglés) cuando nos encontramos con redes neuronales que tienen una capa de entrada (input layer), una o más capas compuestas por perceptrones, llamadas capas ocultas (hidden layers) y una capa final con varios perceptrones llamada la capa de salida (output layer). En general nos referimos a Aprendizaje Profundo cuando el modelo basado en redes neuronales está compuesto por múltiples capas ocultas [Torres18]. En la Figura 3.2 se muestra una MLP sin retroalimentaciones.

Las redes neuronales con múltiples capas son a menudo usadas para clasificación, por ejemplo, reconocimiento de patrones en imágenes, pronóstico de enfermedades, procesamiento natural del lenguaje, reconocimiento de voz, síntesis de voz, predicción de ventas, etc. Este tipo de redes tienen la capacidad de adquirir conocimiento jerarquizado. Esto significa que en las primeras capas de la red se tienen conocimientos muy simples, pero conforme se avanza dentro de la red, el conocimiento adquirido por las capas se torna más

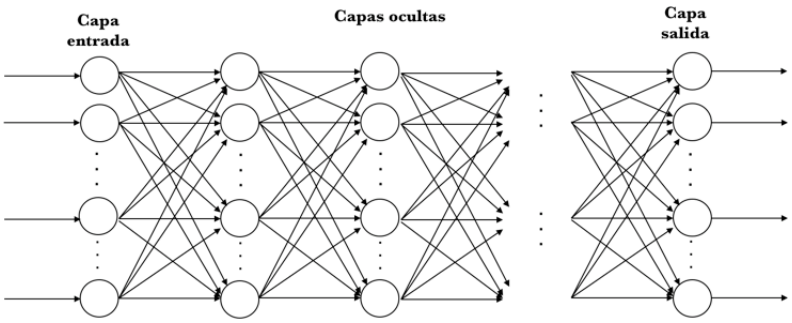

Figura 3.2: Arquitectura de red neuronal multicapa. Fuente [Torres18].

abstracto [Goodfellow16].

3.1.3. Funciones de activación

Una neurona biológica puede estar activa (excitada) o inactiva (no excitada); es decir, que tiene un "estado de activación". En el caso de la neurona artificial la función de activación se encarga de realizar el trabajo de umbral en la neurona biológica, la función activación calcula el estado de actividad de una neurona; transformando la entrada global en un valor (estado) de activación [Caicedo09].

Existen diversas funciones de activación y cada una con características diferentes, por lo que dependiendo del problema a tratar será más conveniente utilizar una u otra función de activación. Algunas de las funciones más usadas en la actualidad son: Lineal, Escalón, Sigmoide, Tangente hiperbólica, Softmax y ReLU [Géron17, Torres18]. En la Figura 3.3 se muestra un ejemplo de algunas funciones de activación.

En esta tesis se utilizaron como funciones de activación: $ReLU$ para las capas ocultas y *Softmax* para la ultima capa, las cuales se describen a continuación:

- ***Softmax***. Es comúnmente utilizada en tareas de clasificación multiclase y se requiere identificar a cuál pertenece la entrada de la red. La función de activación de *Softmax* devuelve la distribución de probabilidad sobre clases de salida mutuamente excluyentes. *Softmax* se encontrará a menudo en la capa de salida de un clasificador [Torres18].

- **Función *ReLU* (unidad rectificada lineal)**. Se trata de una de las funciones de activación más utilizadas. Se comporta como una función constante para valores

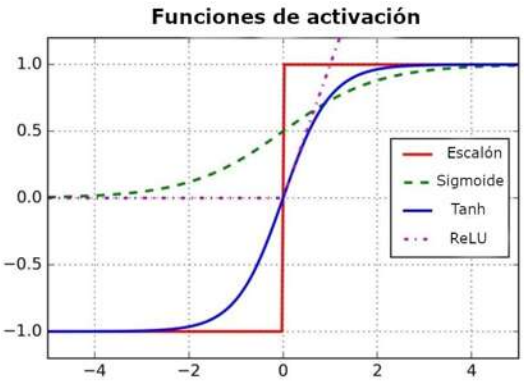

Figura 3.3: Funciones de activación. Fuente [Géron17].

negativos y como una función lineal para valores positivos. En la práctica funciona muy bien y tiene la ventaja de ser rápido de calcular. Lo que es más importante, el hecho de que no tenga un valor de salida máximo también ayuda a que no se sature para grandes valores de entrada.

3.2. Hiperparámetros generales

En general, se considera un parámetro del modelo como una variable de configuración que es interna al modelo y cuyo valor puede ser estimado a partir de los datos. En cambio, por hiperparámetros nos referimos a variables de configuración que son externas al modelo en sí mismo y cuyo valor en general no puede ser estimado a partir de los datos, y son especificados por el programador para ajustar los algoritmos de aprendizaje [Torres18].

3.2.1. Optimizadores

De forma general, podemos ver el proceso de aprendizaje como un problema de optimización global donde los parámetros (los pesos y los sesgos) se deben ajustar de tal manera que la función de costo (*loss*) se minimice. En la mayoría de los casos, estos parámetros no se pueden resolver analíticamente, pero en general se pueden aproximar bien con algoritmos de optimización iterativos u optimizadores [Torres18], como: *SGD*, *RMSprop*, *Adagrad*, *Adadelta*, *Adam*, *Adamax* y *Nadam*. Se puede encontrar más detalle de cada uno de estos enn el siguiente libro [Goodfellow16].

El optimizador utilizado en esta tesis es *Adam* y a continuación se da una breve descripción:

Adam

Es un método de descenso de gradiente estocástico que solo requiere gradientes de primer orden con poco requerimiento de memoria. El método calcula tasas de aprendizaje adaptativas individuales para diferentes parámetros a partir de estimaciones del primer y segundo momento (también conocido como *momentum*, que acelera el descenso) de los gradientes; el nombre *Adam* se deriva de la estimación del momento adaptativo. Este método está diseñado para combinar las ventajas de dos métodos populares: *AdaGrad*, que funciona bien con gradientes escasos, y *RMSProp*, que funciona bien en entornos en línea y no estacionarios; tiene relaciones importantes con estos y otros métodos de optimización estocástica. Algunas de las ventajas de *Adam* son que las magnitudes de las actualizaciones de parámetros son invariantes al reescalado del gradiente, sus pasos están limitados aproximadamente por el hiperparámetro de tamaño de paso, no requiere un parámetro estacionario objetivo, trabaja con gradientes dispersos [Kingma15].

3.2.2. Regularizadores

Existen diversas técnicas de regularización, las mas utilizadas son: regularización L^1, regularización L^2, *Dropout* (Tasa de desactivación), Aumento de Datos, Parada Temprana (*Early Stopping*), regularización de Norma-Máxima (*Max-Norm*) y regularización por Ruido. Sin embargo, solo se mencionaran las técnicas utilizadas en esta tesis. Se puede encontrar más detalle de cada uno de estos en los siguientes libros [Goodfellow16, Géron17, Brownlee19].

Dropout

Podría decirse que la técnica de regularización más popular para las redes neuronales profundas es *dropout*. Es un algoritmo bastante simple: en cada paso de entrenamiento, cada neurona (incluidas las neuronas de entrada, pero excluyendo las de salida) tiene una probabilidad p de ser temporalmente "desactivadas", lo que significa que será ignorada por

completo durante este paso de entrenamiento, pero puede estar activo durante el siguiente paso (ver Figura 3.4). Esto obliga a que las neuronas cercanas no dependan tanto de las neuronas desactivadas. Este método ayuda a reducir el sobre-entrenamiento ya que las neuronas cercanas suelen aprender patrones que se relacionan y estas relaciones pueden llegar a formar un patrón muy especifico con los datos de entrenamiento, con *dropout* esta dependencia entre neuronas es menor en toda la red neuronal, de esta manera la neuronas necesitan trabajar mejor de forma solitaria y no depender tanto de las relaciones con las neuronas vecinas [Géron17].

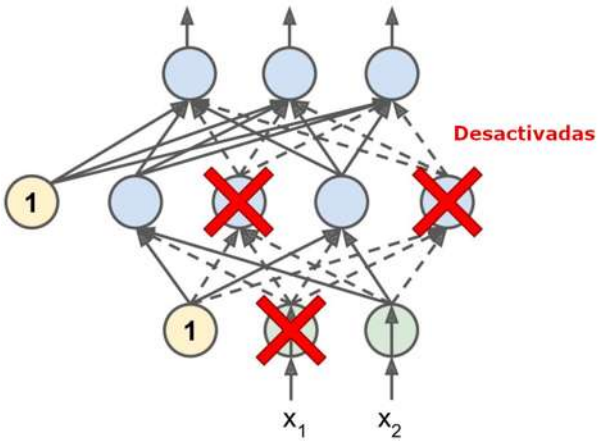

Figura 3.4: Regularizador *dropout*. Fuente [Géron17].

Parada Temprana

Un desafío importante en el entrenamiento de redes neuronales es cuánto tiempo se deben entrenar. Muy poco entrenamiento significará que el modelo no se ajustará bien al conjunto de entrenamiento ni a los conjuntos de prueba. Demasiado entrenamiento significará que el modelo se ajustará demasiado al conjunto de datos de entrenamiento y tendrá un rendimiento deficiente en el conjunto de prueba. Un compromiso es entrenar en el conjunto de datos de entrenamiento, pero detener el entrenamiento en el punto en que el rendimiento en un conjunto de datos de validación comienza a degradarse. Este enfoque simple, efectivo y ampliamente utilizado para entrenar redes neuronales se llama *parada temprana* [Brownlee19]. Consiste en monitorizar el error de validación con la finalidad de

identificar el momento en que comienza a incrementar, lo que indica la presencia del sobre-entrenamiento. Cuando se identifica el sobre-entrenamiento el entrenamiento es detenido y se retornan los parámetros de la red con los que se obtuvo los mejores resultados (ver Figura 3.5). Sin embargo, durante los entrenamientos es probable que el error de validación incremente por instantes para volver a disminuir, por lo que se utiliza el concepto de paciencia. La paciencia consiste en esperar un plazo determinado a que el error disminuya, si el plazo es superado el entrenamiento es detenido [Goodfellow16].

Aunque la *parada temprana* funciona muy bien en la práctica, normalmente puede obtener un rendimiento mucho mayor de su red combinándola con otras técnicas de regularización [Géron17].

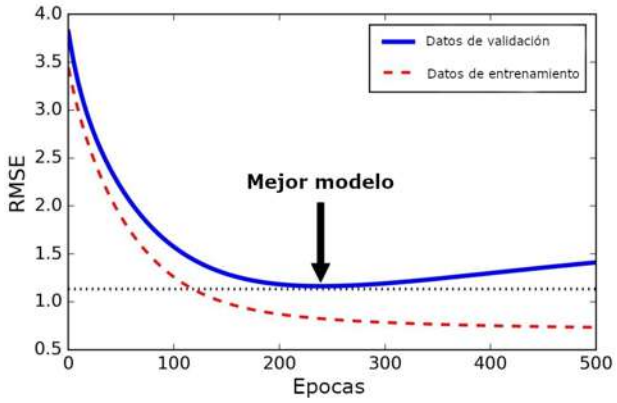

Figura 3.5: Regularizador *Parada Temprana*. Fuente [Géron17].

3.2.3. Decaimiento de la Tasa de Aprendizaje

La mejor *tasa de aprendizaje* en general es aquel que disminuye a medida que el modelo se acerca a una solución. Para conseguir este efecto, este hiperparámetro, es el *decaimiento de la tasa de aprendizaje (learning rate decay)*, que se usa para disminuir la *tasa de aprendizaje* a medida que van pasando épocas para permitir que el aprendizaje avance más rápido al principio con una *tasa de aprendizaje* más grandes. A medida que se avanza, se van haciendo ajustes cada vez más pequeños para facilitar que converja el proceso de entrenamiento al mínimo de la función los [Torres18]. En la Figura 3.6 se muestra una gráfico donde se puede ver que es lo que pasa si la *tasa de aprendizaje* es muy grande,

normal, chico y con decaimiento.

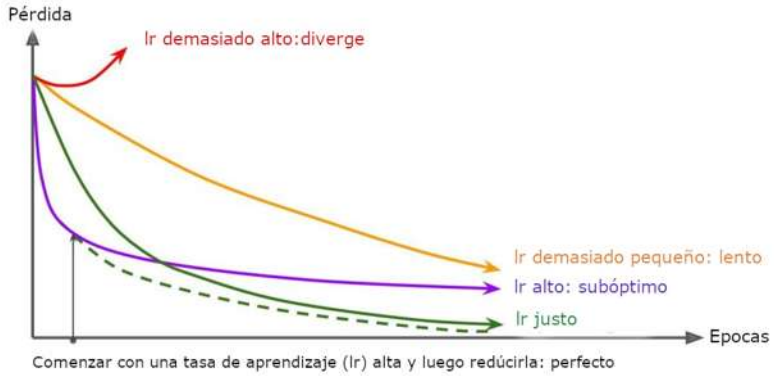

Figura 3.6: Gráfica de como mejora el modelo con *Decaimiento de la Tasa de Aprendizaje*. Fuente [Géron17].

3.2.4. Normalización por Lotes

En un artículo de 2015 [Ioffe15], Sergey Ioffe y Christian Szegedy propusieron una técnica llamada *normalización por lotes* (batch normalization) para abordar los problemas de gradientes que desaparecen o explotan y, de manera más general, el problema de que la distribución de las entradas de cada capa cambia durante el entrenamiento, ya que los parámetros de las capas anteriores cambian.

La técnica consiste en agregar una operación en el modelo justo antes de la función de activación de cada capa, simplemente centrando en cero y normalizando las entradas, luego escalando y cambiando el resultado usando dos nuevos parámetros por capa (uno para escalar, el otro para cambiar). En otras palabras, esta operación permite que el modelo aprenda la escala y la media óptimas de las entradas para cada capa.

Para centrar en cero y normalizar las entradas, el algoritmo necesita estimar la media y la desviación estándar de las entradas. Lo hace evaluando la media y la desviación estándar de las entradas sobre el mini lote actual (de ahí el nombre "*normalización por lotes*").

En el momento de la prueba, no hay un mini lote para calcular la media empírica y la desviación estándar, por lo que simplemente usa la media y la desviación estándar de

todo el conjunto de entrenamiento. Por lo general, estos se calculan de manera eficiente durante el entrenamiento utilizando un promedio móvil.

3.3. Redes Neuronales Convolucionales

Las redes convolucionales, también conocidas como redes neuronales convolucionales, o CNN, son un tipo especializado de red neuronal para procesar datos que tienen una topología tipo cuadrícula. Ya sean unidimensionales, bidimensionales o tridimensionales, las CNN tienen las mismas características y los mismos métodos de procesamiento. La diferencia clave es la dimensionalidad de los datos de entrada y cómo el detector de características (o filtro) se desliza entre los datos [Shenfield20]. Por lo general las CNN 1D son utilizadas en series de tiempo, como podrían ser datos de audio y texto (ver Figura 3.7). Las CNN 2D, regularmente son utilizadas para imágenes y datos de audio que han sido preprocesados con una transformada de Fourier. Y finalemte las CNN 3D, son utilizadas en imágenes médicas, como las tomografías computarizadas y datos de vídeo en color [Goodfellow16].

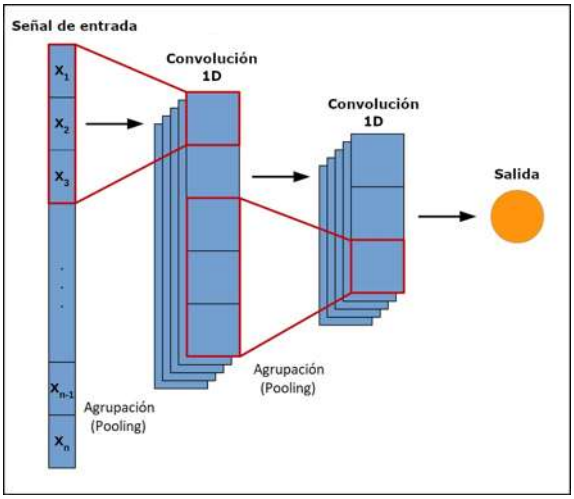

Figura 3.7: Típica arquitectura de una CNN 1D. Fuente [Shenfield20].

El nombre “red neuronal convolucional” indica que la red emplea una operación matemática llamada convolución. La convolución es un tipo especializado de operación lineal. Las redes convolucionales son simplemente redes neuronales que utilizan la convolución en lugar de la multiplicación general de matrices en al menos una de sus capas

[Goodfellow16].

La operación de convolución generalmente se indica con un asterisco y esta dada por la siguiente ecuación:

$$c(t) = (x_1 * x_2)(t), \tag{3.1}$$

donde, el primer argumento (la señal x_1) de la convolución se suele denominar entrada, y el segundo argumento (la señal x_2) es el filtro o kernel de convolución. La salida a veces se denomina mapa de características, ya que relaciona las características extraídas con la posición de la señal en que se encuentran.

Se puede definir a la convolución discreta como:

$$c[n] = x_1[n] * x_2[n] = \sum_{k=-\infty}^{\infty} x_1[k]x_2[n-k]. \tag{3.2}$$

El bloque de construcción más importante de una CNN es la capa convolucional: las neuronas en la primera capa convolucional no están conectadas a cada una de las posiciones del vector de entrada (como lo estaban en las MLP), sino solo a las posiciones de sus campos receptivos. A su vez, cada neurona en la segunda capa convolucional está conectada solo a las neuronas ubicadas dentro de un pequeño rectángulo en la primera capa. Esta arquitectura permite que la red se concentre en características de bajo nivel en la primera capa oculta, luego las ensambla en características de nivel superior en la siguiente capa oculta, y así sucesivamente. Esta estructura jerárquica es común en las imágenes del mundo real, que es una de las razones por las que las CNN funcionan tan bien para el reconocimiento de imágenes [Géron17]. Esta característica tan importante en las capas convolucionales hace posible que aprendan jerarquías espaciales de patrones preservando relaciones espaciales [Torres18].

3.3.1. Filtrado o kernel de convolución

Los pesos de una neurona se pueden representar como un pequeño vector del tamaño del campo receptivo [Géron17]. En una capa convolucional el filtro se aplica a todo el vector, por lo que la operación de convolución se aplica reiteradas veces desplazando

el filtro. El desplazamiento se suele hacer moviendo el filtro por todas las posiciones del vector, aunque esto puede ser modificado dando pasos de tamaño distinto [Goodfellow16]. Durante el entrenamiento, una CNN encuentra los filtros más útiles para su tarea y aprende a combinarlos en patrones más complejos [Géron17].

3.3.2. Tamaño de paso

En las CNN se pueden usar diferentes longitudes de pasos de avance para la ventada de los filtros. Valores de *tamaño de paso* grandes hacen decrecer el tamaño de la información que se pasará la siguiente capa, también conocido como *stride* [Torres18].

3.3.3. Relleno de ceros

En las redes neuronales convolucionales también se puede aplicar una técnica de relleno de ceros alrededor del margen del vector para mejorar el barrido que se realiza con la ventana del filtro que se va deslizando y mantener el mismo tamaño del vector original. El parámetro para definir este relleno recibe el nombre de *padding*. Los tipos de *relleno de ceros* más utilizados son *relleno-igual* (same-padding) y *relleno-válido* (valid padding). *Relleno-igual* nos permite tener mapas de características del mismo tamaño que el vector original, se añaden tantos ceros como sea necesario para que la salida tenga el mismo tamaño que la entrada. Mientras que *relleno-válido*, indica no hacer *relleno*, es decir, no agregar ceros alrededor del margen del vector y el resultado es una mapa de características de menor tamaño [Torres18].

3.3.4. Capas de agrupación

Además de las capas convolucionales que se describieron anteriormente, las CNN suelen acompañar a la capa de convolución con unas capas de reducción, también conocidas como *pooling*, se aplican inmediatamente después de las capas convolucionales. Una primera aproximación para entender para qué sirven estas capas es ver que las capas de *agrupación* hacen una simplificación de la información recogida por la capa convolucional y crean una versión condensada de la información contenida en estas [Torres18]. La operación de *agru-*

pación en realidad es un submuestreo que ayuda a obtener la información más relevante de los mapas de características [Goodfellow16].

La *agrupación* puede ser útil si nos importa más si alguna característica está presente que dónde está exactamente. Por ejemplo, al determinar si una imagen contiene una cara, no necesitamos saber la ubicación de los ojos con una precisión perfecta de píxeles, solo necesitamos saber que hay un ojo en el lado izquierdo de la cara y un ojo en el lado derecho. lado de la cara. En otros contextos, es más importante preservar la ubicación de una característica. Por ejemplo, si queremos encontrar una esquina definida por dos bordes que se encuentran en una orientación específica, debemos conservar la ubicación de los bordes lo suficientemente bien como para probar si se encuentran [Goodfellow16].

Existen dos formas típicas de realizar el submuestreo, *agrupación-máxima* (max-pooling) y *agrupación-promedio* (average-pooling). *Agrupación-máxima*, se queda con el valor máximo de los valores de la ventana analizada. Mientas que *agrupación-promedio*, se queda con el promedio de los valores de la ventana analizada [Torres18]. Además, existe una variante de la *agrupación* y es *agrupación-global* (global-pooling).

Agrupación-Global

Existen dos tipos de *agrupación-global*: *agrupación-máxima-global* y *agrupación-promedio-global*. *Agrupación-promedio-global*, lo que hace es tomar un promedio de cada mapa de características entrante. Mientras que *agrupación-máxima-global* toma el máximo de cada mapa de características entrante [Brunel19]. En la Figura 3.8 se muestra un esquema del mecanismo de *agrupación-máxima-global*.

Las CNN convencionales realizan la convolución en las capas inferiores de la red, para la clasificación, los mapas de características de la última capa convolucional se vectorizan y se introducen en capas densamente conectadas seguidas de una capa de regresión logística *softmax*. Esta estructura une la estructura convolucional con los clasificadores de redes neuronales tradicionales. Trata las capas convolucionales como extractores de características y la característica resultante se clasifica de forma tradicional [Lin14].

Con *agrupación-global*, en lugar de agregar capas densamente conectadas sobre los mapas de características, se toma el máximo o el promedio (según sea el caso) de cada mapa

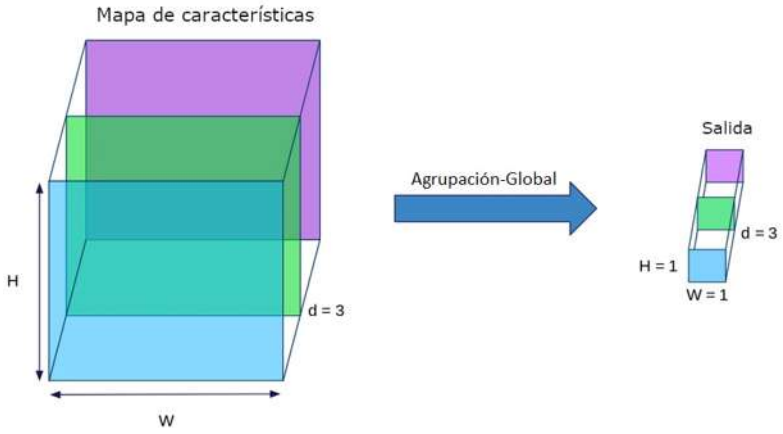

Figura 3.8: Esquema del mecanismo de *agrupación-global*. *Agrupación-global* realiza una operación que toma el máximo o el promedio por mapa de características y produce un vector 1D con un tamaño conocido (número de mapas de características). Fuente [Brunel19].

de características y el vector resultante se introduce directamente en la capa *softmax*. Una ventaja de *agrupación-global* sobre las capas densamente conectadas es que es más nativo a la estructura de convolución al hacer cumplir las correspondencias entre los mapas de características y las categorías. Otra ventaja es que no hay ningún parámetro para optimizar en *agrupación-global*, por lo que se evita el sobre-entrenamiento en esta capa. Además, *agrupación-global* suma la información espacial, por lo que es más resistente a fluctuaciones en las entradas [Lin14]. En la Figura 3.9 se muestra un ejemplo de las diferencias entre la última capa densamente conectada y el uso de *agrupación-promedio-global*.

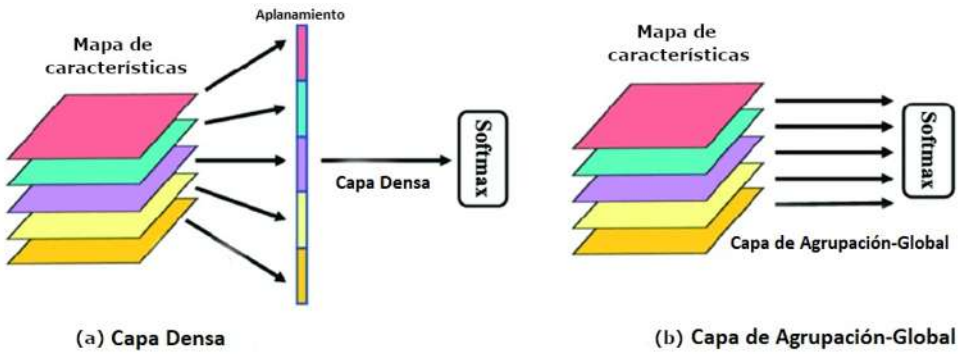

Figura 3.9: Comparación de la capa densamente conectada y la capa de *agrupación-promedio-global*. Fuente [Zhang21].

3.4. Sub-entrenamiento y Sobre-entrenamiento

El objetivo de una red neuronal es tener un modelo final que funcione bien tanto con los datos que usamos para entrenarlo como con los nuevos datos en los que se usará el modelo para hacer predicciones. Aprender y también generalizar a nuevos casos es difícil. Muy poco aprendizaje y el modelo tendrá un rendimiento deficiente en el conjunto de datos de entrenamiento y en los datos nuevos. El modelo no se ajustará al problema. Demasiado aprendizaje y el modelo funcionará bien en el conjunto de datos de entrenamiento y mal en datos nuevos, el modelo se ajustará demasiado al problema. En ambos casos, el modelo no se ha generalizado, en la Figura 3.10 se muestran las tres representaciones de los casos antes descritos [Brownlee19].

- **Modelo sub-entrenado**. Un modelo que no logra aprender suficientemente el problema y se desempeña mal en un conjunto de datos de entrenamiento y no funciona bien en una muestra reservada.

- **Modelo sobre-entrenado**. Un modelo que aprende demasiado bien el conjunto de datos de entrenamiento, se desempeña bien en el conjunto de datos de entrenamiento, pero no funciona bien en una muestra reservada.

- **Modelo bien-entrenado**.Un modelo que aprende adecuadamente el conjunto de datos de entrenamiento y generaliza bien al conjunto de datos de reserva.

Podemos abordar el sub-entrenamiento aumentando la capacidad del modelo. La capacidad se refiere a la habilidad de un modelo para adaptarse a una variedad de funciones; más capacidad, significa que un modelo puede adaptarse a más tipos de funciones para mapear entradas a salidas. El aumento de la capacidad de un modelo se logra fácilmente cambiando la estructura del modelo, como agregar más capas y/o más nodos a las capas. Debido a que un modelo de ajuste insuficiente se aborda tan fácilmente, es más común tener un modelo de ajuste excesivo [Brownlee19].

Un modelo sobre-entrenado se diagnostica fácilmente monitoreando el rendimiento del modelo durante el entrenamiento evaluándolo tanto en un conjunto de datos de entrenamiento como en un conjunto de datos de validación de reserva. Los gráficos de líneas

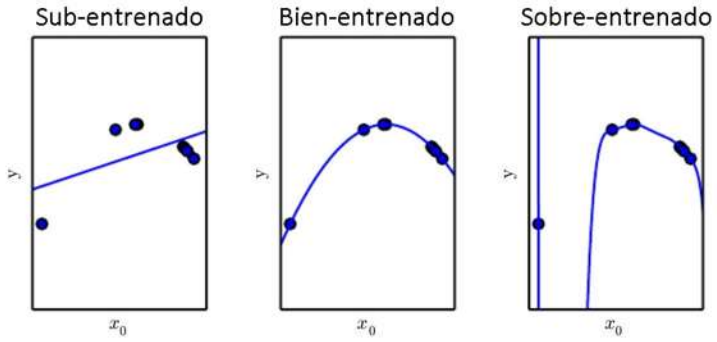

Figura 3.10: Problemas de entrenamiento del modelo. Figura de la izquierda, el modelo presenta sub-entrenamiento, ya que no es capas de separar correctamente los datos. Figura del centro, el modelo generaliza bien. Figura de la derecha, presenta sobre-entrenamiento, puesto que genera una frontera muy compleja. Fuente [Goodfellow16].

del rendimiento del modelo durante el entrenamiento, llamada función de pérdida (*loss*), mostrarán un patrón familiar [Brownlee19]. En la Figura 3.11 se muestra un ejemplo de sobre-entrenamiento.

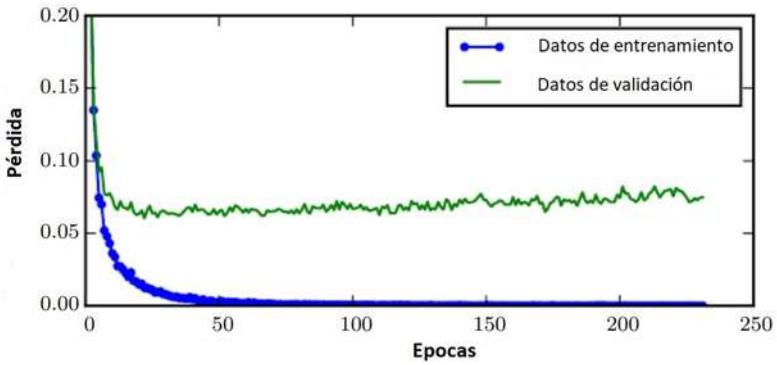

Figura 3.11: Ejemplo de sobre-entrenamiento. Fuente [Goodfellow16].

Dos formas de abordar un modelo sobre-entrenado:

1. Entrenando la red con más ejemplos.

2. Cambiando la complejidad de la red.

Una ventaja de las redes neuronales muy profundas es que su rendimiento continúa mejorando a medida que se alimentan de conjuntos de datos cada vez más grandes. Un

modelo con un número casi infinito de ejemplos eventualmente se estancará en términos de lo que la capacidad de la red es capaz de aprender. Un modelo puede sobre-entrenar un conjunto de datos de entrenamiento porque tiene la capacidad suficiente para hacerlo. Reducir la capacidad del modelo reduce la probabilidad de que el modelo sobre-entrene el conjunto de datos de entrenamiento, hasta un punto en el que ya no sobre-entrene. La capacidad de un modelo de red neuronal, su complejidad, se define tanto por su estructura en términos de nodos y capas como por los parámetros en términos de sus pesos [Brownlee19].

Las técnicas para reducir el sobre-entrenamiento utilizadas en esta tesis son: *dropout*, *parada temprana*, *decaimiento de la tasa de aprendizaje*, *normalización por lotes* y *agrupación-máxima-global*.

3.5. Arquitectura de la Red Neuronal Convolucional utilizada

Comenzamos con un modelo con solo unas pocas capas convolucionales y un número limitado de filtros por capa convolucional, después aumentamos el número de capas y filtros, variamos el tamaño del kernel de convolución, el tipo de *agrupación* y el tamaño de su ventana, al final nos quedamos con el modelo CNN que mejor clasificaba los datos. Nuestra CNN 1D tiene 12 capas convolucionales; la ultima capa utiliza la función de activación *Softmax* y con un número de filtros igual al número de parlantes, en todas las demás capas se utilizan la función de activación *ReLU*.

En la arquitectura de red neuronal convolucional clásica, las primeras capas, las capas convolucionales, actúan como extractores de características, mientras que las últimas capas, normalmente capas densas, conforman el clasificador real, también el número de neuronas de la última capa densa normalmente es igual al número de clases y la función de activación *Softmax* se utiliza en esta última capa. Las redes neuronales convolucionales convencionales asumen que las clases son linealmente separables, lo cual es una limitación, además las capas densas son propensas al sobre-entrenamiento. Por estas razones, preferimos utilizar la estrategia de *agrupación-máxima-global*, que reemplaza las capas densas con capas convolucionales adicionales que conforman una estructura de "micro red"que actúa como

un aproximador de función no lineal que genera un mapa de características para cada clase en la última capa, donde *agrupación-máxima-global* se utiliza para la selección de clases. *Agrupación-máxima-global* impone correspondencias entre mapas de características y categorías, siendo más natural para la estructura de convolución que las capas densas y es más resistente a los cambios de tiempo en la señal de entrada.

En la Tabla 3.1 se muestra el resumen del modelo, en esta se puede ver la cantidad de parámetros utilizados, el tipo de capas, así como las dimensiones de la entrada y salida en cada capa de la red. La última capa de esta tabla tiene una cantidad de filtros que equivale a la cantidad de parlantes del conjunto de datos, por ejemplo, en el caso de la base de datos ELSDSR hay 22 parlantes, por lo que el número de filtros de la última capa es 22. Para reducir aún más el sobre-entrenamiento, utilizamos: un *dropout* = 0.25, una paciencia para la *parada temprana* = 30, una paciencia para el *decaimiento de la tasa de aprendizaje* = 20 y *normalización por lotes* (ya que este actúa también como regularizador).

3.6. Conclusiones del capítulo

Las CNN 1D resultan ser muy buenas para la clasificación de series de tiempo, ya que estas son muy eficaces para extraer las características principales de los datos de entrada, preservando las posiciones espaciales de estos. Una ventaja de las redes neuronales muy profundas es que su rendimiento continúa mejorando a medida que se alimentan de conjuntos de datos cada vez más grandes. Sin embargo, las redes neuronales convolucionales contienen muchos hiperparámetros y variables para su diseño, como el número de capas, tamaño de kernel, función de activación, tasa de aprendizaje, optimizador, regularizadores, etc., lo que las vuelve complejas de diseñar.

Tabla 3.1: Especificaciones para la CNN 1D utilizada para la base de datos ELSDSR con 22 parlantes. ks=tamaño del kernel; ws=tamaño de la ventana; ds=tasa de desactivación.

Tipo de capa	Dimensión del tensor	# Parámetro
Capa de entrada	(480, 1)	0
Conv1d_1 ks=3, filtros=32	(478, 32)	128
Conv1d_2 ks=3, filtros=64	(476, 64)	6,208
Dropout_1 ds=0.25	(476, 64)	0
Conv1d_3 ks=7, filtros=32	(470, 32)	14,368
Conv1d_4 ks=7, filtros=64	(464, 64)	14,400
Conv1d_5 ks=7, filtros=64	(458, 64)	28,736
Max_pooling1d ws=2	(229, 64)	0
Dropout_2 ds=0.25	(229, 64)	0
Conv1d_6 ks=20, filtros=32	(210, 32)	40,992
Conv1d_7 ks=20, filtros=64	(191, 64)	41,024
Conv1d_8 ks=20, filtros=96	(172, 96)	122,976
Dropout_3 ds=0.25	(172, 96)	0
Conv1d_9 ks=30, filtros=64	(143, 64)	184,384
Conv1d_10 ks=30, filtros=96	(114, 96)	184,416
Dropout_4 ds=0.25	(114, 96)	0
Conv1d_11 ks=30, filtros=128	(85, 128)	368,768
Dropout_5 ds=0.25	(85, 128)	0
Conv1d_12 ks=30, filtros=22 (filtros = número de parlantes)	(56, 22)	84,502
Global_max_pooling1d	(22)	0
		Total: 1,090,902

Capítulo 4

Implementación

En este Capítulo se describen las etapas que se llevaron a cabo para la estimación del pulso glotal, así como las características y aspectos mas importantes del sistema, durante su diseño, además, de hablar sobre las bases de datos utilizadas.

4.1. Bases de datos utilizadas

Las bases de datos utilizadas en esta tesis para los experimentos consta de dos corpus que son: Base de Datos de Voz en Inglés para el Reconocimiento de Parlantes (ELSDSR, por sus siglas en inglés) y Corpus de Voz Continua Acústica-Fonética (TIMIT, por las siglas de las instituciones que la crearón).

4.1.1. ELSDSR

El corpus ELSDSR consta de grabaciones de 20 daneses, un islandés y un canadiense, 10 mujeres y 12 hombres, con edades comprendidas entre los 24 y los 63 años. Este es el trabajo de la facultad, los estudiantes de doctorado y los estudiantes de maestría del departamento de informática y modelado matemático (IMM) de la Universidad Técnica de Dinamarca, se conoce como la base de datos de voz en inglés para el reconocimiento de parlantes (ELSDSR) [Feng05]. ELSDSR se divide en conjuntos de prueba y entrenamiento. El texto de entrenamiento tiene 7 párrafos, los 22 parlantes leen cada párrafo, por lo que hay $22 \times 7 = 154$ grabaciones. Para el conjunto de prueba, cada parlante leyó dos oraciones

diferentes de un texto sobre las pirámides egipcias disponible en [Home22], por lo que consta de $22 \times 2 = 44$ grabaciones.

4.1.2. TIMIT

El corpus TIMIT está diseñado para proporcionar datos de voz para estudios acústico-fonéticos y para el desarrollo y evaluación de sistemas automáticos de reconocimiento de voz. TIMIT contiene grabaciones de 630 parlantes (438 hombres y 192 mujeres) de ocho dialectos principales del inglés americano, cada uno leyendo diez oraciones fonéticamente ricas. El corpus TIMIT incluye archivos de audio de voz de 16 bits y muestreados a 16 kHz para cada expresión, la duración de los archivos de voz varía aproximadamente de 3s a 4.5s. El diseño del corpus fue un esfuerzo conjunto entre el Instituto de Tecnología de Massachusetts (MIT, por sus siglas en inglés), SRI International (SRI) y Instrumentos Texas, Inc. (TI, por sus siglas en inglés). El discurso fue grabado en TI, transcrito en MIT y verificado y preparado para producción en CD-ROM por el Instituto Nacional de Estándares y Tecnología (NIST, por sus siglas en inglés). Para mayor información sobre esta base de datos dirigirse a [TI22]. Este corpus se divide en conjuntos de prueba y entrenamiento; para cada parlante, se usan 8 archivos de voz para entrenamiento ($630 \times 8 = 5,040$ archivos de voz) y 2 archivos de voz son utilizados para prueba ($630 \times 2 = 1,260$ archivos de voz).

4.1.3. Conjuntos de entrenamiento, prueba y validación

El conjunto de datos de entrenamiento es utilizado para extraer los trenes de pulsos glotales de la voz y posteriormente el modelo CNN se utiliza para que aprenda los patrones de la señal. Mientras que el conjunto de prueba es utilizado para evaluar el comportamiento del modelo CNN, ante señales distintas a las utilizadas para entrenar, por lo general, se utiliza un conjunto de datos de validación que es utilizado para monitorear a la CNN durante su entrenamiento (identificar cuando se presenta sobre-entrenamiento), este conjunto de datos se suele extraer del conjunto de datos de entrenamiento. Para esta tesis se destina el 20 % del conjunto de tren de pulsos glotales de entrenamiento para el conjunto de datos de validación. En la Figura 4.1 se muestra un diagrama donde se resume la distribución de los datos.

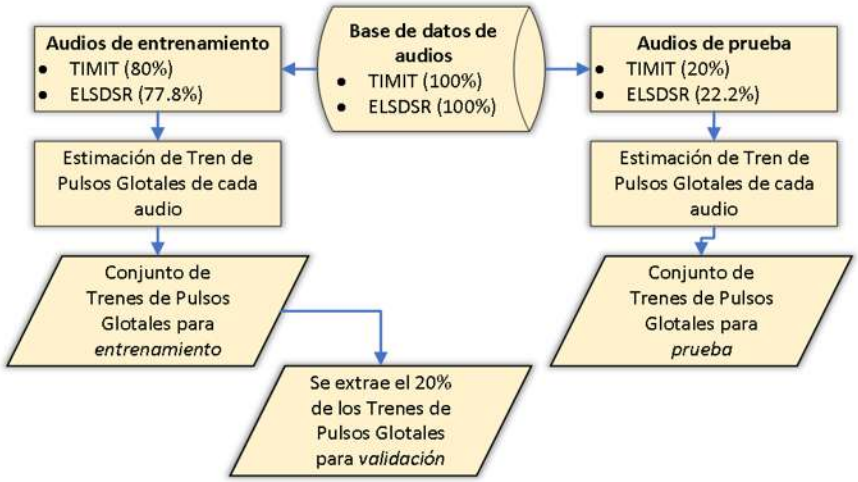

Figura 4.1: Distribución de los conjuntos de datos.

4.2. Estimación del Pulso Glotal

La implementación del sistema se lleva a cabo en el lenguaje de programación Python, ya que este cuenta con una gran cantidad de bibliotecas (en específico de inteligencia artificial, como Keras de Tensor Flow), marcos, extensiones de archivo y colecciones de módulos. Por último, es de código abierto y tiene a sus espaldas a una gran comunidad de desarrolladores que siempre están mejorándolo y perfeccionándolo.

Para esta tesis proponemos un método de identificación texto-independiente de parlantes por medio del pulso glotal y de un clasificador CNN 1D. En la Figura 4.2 se muestra el diagrama de flujo del proceso en general que utilizamos para la estimación de los pulsos glotales para cada audio. Este comienza leyendo el archivo de audio; la señal de voz obtenida se discrimina entre silencio y voz, luego, se divide en marcos de 30 ms con traslape de dos tercios entre los marcos consecutivos, cada marco es examinado para conocer si contiene voz, si contiene voz se analiza el marco para saber si contiene sonido vocalizado y se les estima su tren de pulsos glotales.

4.2.1. Discriminador entre silencio y voz

En el modelo que presentamos, se utiliza la energía de tiempo corto y el régimen de cruces por cero de tiempo corto para identificar el inicio y el fin de la señal de voz

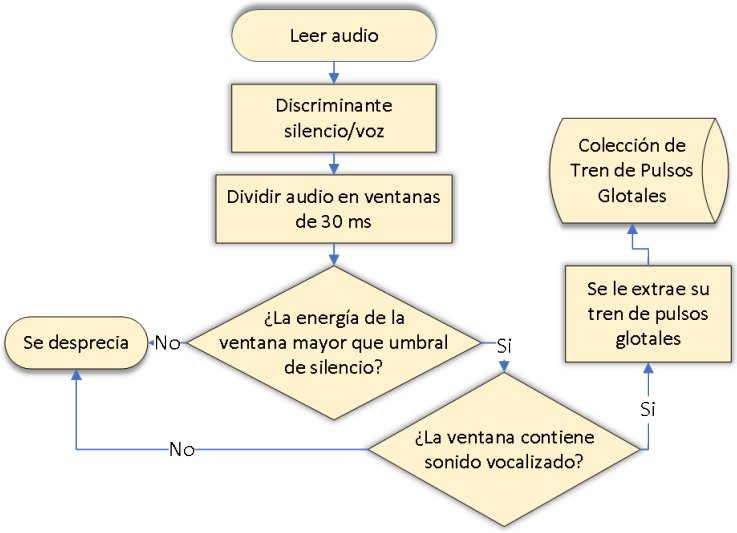

Figura 4.2: Diagrama de estimación de los pulsos glotales.

(se elimina el silencio o ruido de fondo). Comúnmente se suelen utilizar en conjunto como discriminador entre silencio y voz; la energía de tiempo corto y el régimen de cruces por cero, ya que cada técnica tiene ciertas cualidades una sobre la otra, por ejemplo el régimen de cruces por cero tiempo corto detecta mejor los sonidos fricativos y plosivos (oclusivos). En cambio la energía de tiempo corto no es muy eficaz para detectarlos ya que estos sonidos contienen baja energía. Sin embargo, la energía de tiempo corto detecta muy bien los sonidos vocalizados, ya que son los sonidos con mayor energía. En la Figura 4.3 se expone un ejemplo del recorte de una señal de voz, donde se eliminan los extremos de la señal que contenían silencio o ruido de fondo.

Es importante conservar algunos de los sonidos que acompañan los sonidos vocalizados, ya que como se mencionó anteriormente, la voz depende en gran medida del contexto en el que se producen los sonidos; es decir, los sonidos que ocurren antes y después del sonido actual (coarticulación). Es por esto que utilizamos estas dos técnicas para el discriminador silencio/voz.

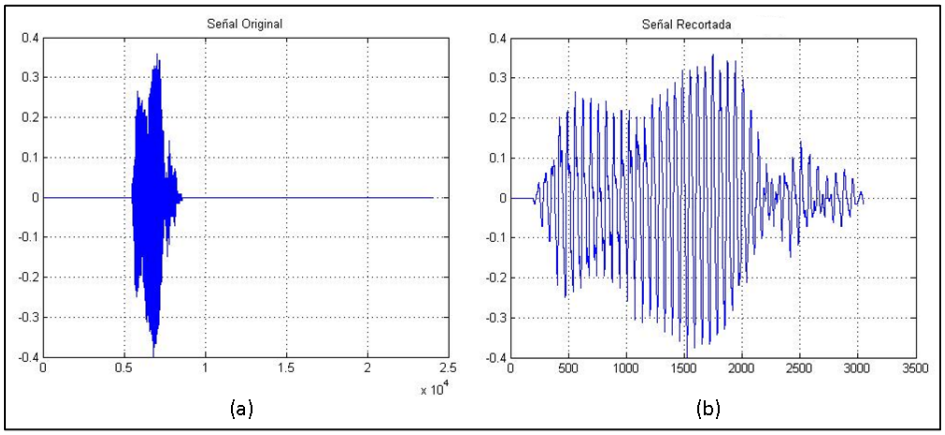

Figura 4.3: Ejemplo discriminador silencio/voz. Señal del audio (a) y señal recortada (b).

4.2.2. División de audio en segmentos

Se divide el audio en ventanas o segmentos, ya que como se mencionó antes, en el marco de la TFF, el pulso glotal siempre esta presente en los sonidos vocalizados. Las propiedades de estos sonidos persisten por un tiempo apreciable o cambian muy lentamente mientras se mantenga la configuración del tracto vocal. El filtro del tracto vocal, suele considerarse lineal e invariante en el tiempo, en segmentos de señal de voz de corta duración. El procedimiento se lleva acabo utilizando una ventana de 30 ms (480 elementos, ya que $f_s = 16$ KHz) que se desplaza 10 ms (160 elementos) entre cada ventana, esto permite que la ventana actual comparta información con las dos ventanas anteriores y con las dos ventanas posteriores. La ventana es desplazada por todo el audio, obteniendo como resultado una lista de segmentos. En nuestro modelo se decidió utilizar la ventana rectangular, ya que es importante conservar la forma de onda de la señal.

4.2.3. Identificación de segmentos vocalizadas

Antes de realizar el análisis de identificación de segmentos vocalizados/no vocalizados; se calcula la energía del segmento analizado y se compara con umbral de silencio (en esta tesis es de 10^{-4}), para conocer si este contiene voz, y si no es así no realizar operaciones innecesarias.

Si el segmento contiene voz pasa por un proceso para identificar aquellos que tienen

sonido vocalizado. En esta tesis se utilizo el error de predicción y la autocorrelación normalizada de este. Se utiliza la autocorrelación ya que esta amplifica las muestras de la señal que son periódicas y las que no las atenúa, facilitando la detección de los sonidos vocalizados. Se elije como sonido vocalizado cuando el nivel del pico más grande de autocorrelación normalizada esta por encima de un valor umbral, en esta tesis es de 0.2.

En la Figura 4.4 se expone un ejemplo de un segmento que contiene sonido vocalizado, donde se puede observar que se elimina el primer pico de la autocorrelación que corresponde a la energía de la ventana y de esta manera identificar de una manera mas fácil el valor del segundo pico más grande, en promedio el rango del tono es de 125 Hz para el hombre, 250 Hz para la mujer y 350 Hz en la infancia.

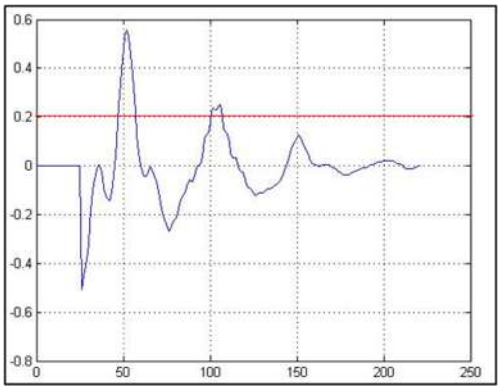

Figura 4.4: Ejemplo de identificación de sonido vocalizado.

4.2.4. Método propio para estimar el Pulso Glotal

Nuestro método para extraer el pulso glotal es iterativo (ver Sección 2.4.2), similar a las técnicas de Murphy [Murphy08] o IAIF de Alku [Alku92a].

En el Algoritmo 2 se presenta nuestro método para la estimación del tren de pulsos glotales para un marco (cada marco tiene una longitud de 480 elementos). Nuestro sistema comienza por calcular los coeficientes LPC α_p del marco (que son también los coeficientes que modelan el tracto vocal), con un orden del predictor de $p = 18$, luego, se aplica filtrado inverso a la señal de voz $s^{(i)}[n]$ (donde $i = 0$) y se obtiene una primera estimación del pulso glotal $u^{(i)}[n]$. Posteriormente, se desglotaliza la señal de voz, es decir, a la señal de voz se

le resta la primera estimación del pulso glotal y reemplazamos la señal de voz $s^{(i)}[n]$ por $s^{(i+1)}[n]$ y repetimos el análisis LPC; filtrado inverso; y desglotalización, hasta un máximo de iteraciones o hasta que la señal de voz desglotalizada cambie demasiado poco. Finalmente, a la señal original $s^{(0)}[n]$ se le aplica filtrado inverso con los coeficientes del modelo del tracto vocal que se calculo de manera iterativa α_p y se obtiene el pulso glotal definitivo.

Algoritmo 2 Método para estimar el pulso glotal

Entrada: $s^{(0)}[n]$, p
Salida: $u^{(i)}[n]$
$i \leftarrow 0$;
repetir
  $\alpha_p \leftarrow LPC(s^{(i)}[n], p)$; `/* Coeficientes del tracto vocal */`
  $u^{(i)}[n] \leftarrow FiltroInverso(\alpha_p, 1, s^{(i)}[n])$; `/* Estimación del pulso glotal */`
  $s^{(i+1)}[n] \leftarrow s^{(i)}[n] - u^{(i)}[n]$; `/* Desglotalizado de la señal de voz */`
  $i \leftarrow i + 1$
hasta que $||s^{(i)}[n] - s^{(i-1)}[n]|| < Tol$;
$u^{(i)}[n] \leftarrow FiltroInverso(\alpha_p, 1, s^{(0)}[n])$; `/* Estimación final del pulso glotal */`
devolver $u^{(i)}[n]$

En la Figura 4.5(a) se muestra un segmento de voz con sonido vocalizado, este segmento en particular corresponde a la vocal /a/ y es el tipo de señal que sería la entrada del Algoritmo 2, la salida de este procedimiento para la misma señal se muestra en la Figura 4.5(b).

4.2.5. Colección de Tren de Pulsos Glotales

Cada parlante tiene relacionado un conjunto de audios divididos en conjunto de prueba y de entrenamiento, como se describió en la Sección 4.1, por lo tanto, cada parlante tiene una colección de trenes de pulsos glotales para el entrenamiento y otra para prueba.

Existe una consideración extra para la evaluación con los audios de prueba, ya que la CNN se va entrenar con marcos fijos de longitud de 480 elementos, y los audios no contienen el mismo número de marcos de trenes de pulsos glotales, por lo que después de ser clasificado cada marco por la CNN, se implementa un esquema de votación, donde cada marco vota por un parlante y el parlante con más votos gana. Para el esquema de votación se utiliza la técnica “moda estadística”, ya que la moda es aquel valor que, dentro

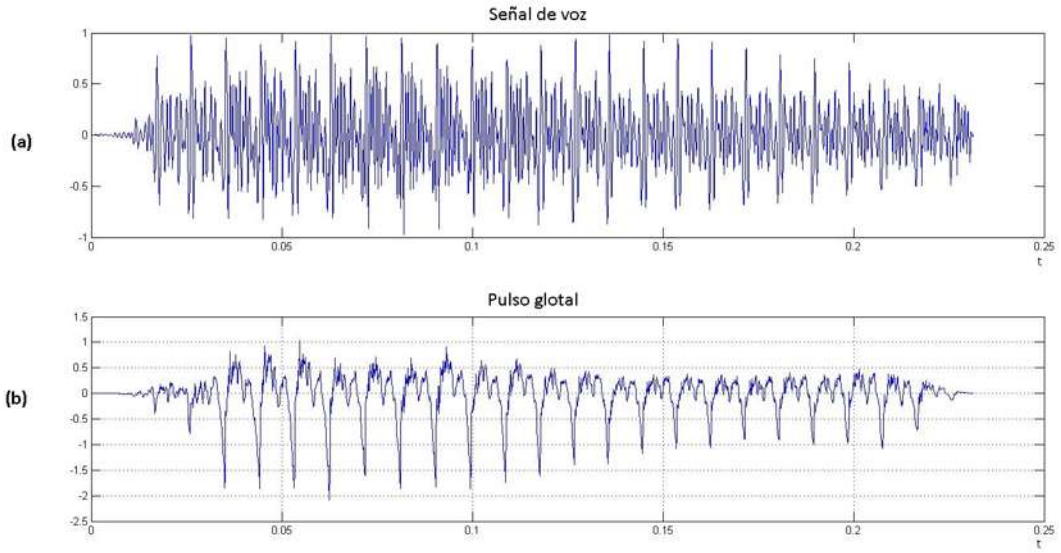

Figura 4.5: Pulso glotal de un segmento con sonido vocalizado. (a), señal de voz de la vocal /a/; (b), tren de pulsos glotales correspondiente a la vocal /a/.

de un conjunto de datos, se repite el mayor número de veces. En la Figura 4.6 se describe el proceso que se lleva acabo para cada audio de prueba, donde se muestra que para cada audio de prueba se encuentran los marcos vocalizados, a cada uno de estos se le estima su tren de pulsos glotales y este es clasificado por la CNN, cada marco vota por un parlante y el parlante con más votos es el parlante ganador para ese audio.

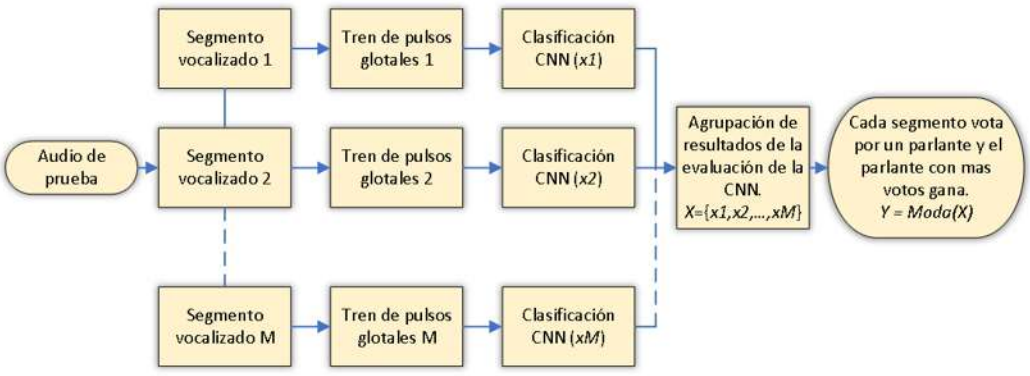

Figura 4.6: Proceso de clasificación de cada audio de prueba.

4.3. Conclusiones del capítulo

La identificación de parlantes utilizando el pulso glotal y CNN 1D, es un proceso complejo ya que para poder estimar los trenes de pulsos glotales, antes debemos realizar un preprocesamiento de la señal de voz y esto incluye una serie de técnicas como: discriminador silencio/voz, segmentación de la señal y detección de segmentos vocalizados. Además de necesitar una CNN 1D para la clasificación de estos trenes. La estimación del pulso glotal de la voz por medio del método propuesto, resulta ser muy exacta, ya que este itera hasta encontrar una estimación estable del tracto vocal, para luego eliminar de la señal de voz y obtener una buena estimación del pulso glotal, con un menor costo computacional comparado con otras técnicas. Además de que no se necesita de una detección del GCI y sincronización con el tono.

Capítulo 5

Resultados

Es este Capítulo se describen los experimentos realizados, así como las características y aspectos mas importantes para la identificación de parlantes utilizando el pulso glotal y CNN 1D. Además de las propiedades principales de la CNN 1D durante su entrenamiento y para la evaluación de los datos de prueba, También se describen las métricas utilizadas para medir el desempeño del sistema. Finalmente, se exponen los resultados obtenidos por el modelo con las bases de datos utilizadas y se realiza una comparación con otros sistemas de reconocimiento de parlantes.

5.1. Métricas utilizadas

La matriz de confusión se suele utilizar para observar el desempeño en algoritmos entrenados por medio de aprendizaje supervisado. En las matrices de confusión los renglones representan a las clases reales mientras que las columnas representan a las predicciones. Las matrices de confusión en sí mismo no son una medida de desempeño como tal, pero nos permiten observar cuales son las clases que está confundiendo la red neuronal, ya que en la matriz sus celdas almacenan la cantidad de datos clasificados de acuerdo a la predicción y la clase real. Esto permite que al observar los valores en la diagonal de la matriz se pueda observar la cantidad de datos que han sido clasificados correctamente por cada clase. La mayoría de métricas se pueden derivar de la definición de la matriz de confusión. En la Figura 5.1 se muestran las partes principales que conforman a una matriz de confusión,

donde:

TP, son los positivos verdaderos,

TN, son los falsos verdaderos,

FP, son los positivos falsos,

FN, son los negativos falsos.

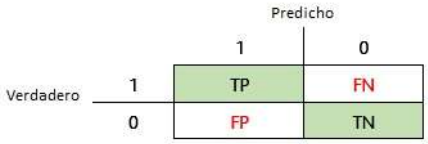

Figura 5.1: Partes de una matriz de confusión binaria.

La Tasa de Verdaderos Positivos (TPR), también conocida como exhaustividad (recall) o sensibilidad, se define como el número de clases que el sistema identifica correctamente (es decir TP, los positivos verdaderos) dividido por el número de clases que el sistema debería tener (es decir TP + FN, positivos). TPR se calcula con la Ecuación 5.1.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5.1}$$

La Tasa de Falsos Positivos (FPR), también conocida como 1 - especificidad, mide la frecuencia con la que el sistema confunde a una clase con otra. FPR se calcula con la Ecuación 5.2.

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{5.2}$$

La precisión (*accuracy*) mide el porcentaje de casos que el modelo ha acertado. Esta es una de las métricas más usadas y favoritas. Sin embargo, la métrica *precisión* puede resultar confusa cuando las clases están desbalanceadas (cuando las clases no tienen el mismo numero de ejemplos). El *precisión* se calcula con la Ecuación 5.3

$$precisión = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{5.3}$$

La curva Característica Operativa del Receptor (ROC, por sus siglas en inglés), es un gráfico que ilustra la capacidad de diagnóstico de un sistema a medida que varia su umbral de descriminación. El área bajo una curva (AUC, por sus siglas en inglés), es el porcentaje del área que esta bajo la curva ROC, cuanto mayor sea esta mejor se considera que es el clasificador. En el plano de la curva ROC, los ejes vertical y horizontal corresponden a TPR y FPR respectivamente. La curva ROC por lo general se utiliza solo para problemas de dos clases, pero es posible realizarla también para problemas multiclase, se conoce como el enfoque de uno contra todos, en el siguiente enlace [scikitlearndevelopers22] se encuentra una explicación mas detallada y ejemplos de como realizarla.

Existen dos métricas ligeramente distintas para las curvas ROC, micro y macro promedio. Un macro-promedio calcula la métrica de forma independiente para cada clase y luego toma la media (trata todas las clases por igual), mientras que un micro-promedio agrega las contribuciones de todas las clases para calcular la media de la métrica. En una configuración de clasificación con varias categorías, micro-promedio es preferible si se sospecha que puede haber un desequilibrio entre las clases [scikitlearndevelopers22].

5.2. Experimentos

Los primeros experimentos realizados es sobre los pulsos glotales, para conocer que tan similares o diferentes son los pulsos glotales (estimados por nuestro método) de distintos sonidos vocalizados, dichos por el mismo parlante y también que tan diferentes son estos comparado con otros parlantes. En la Figura 5.2 se muestra el pulso glotal de las vocales /a/,/e/,/i/,/o/ y /u/, pronunciadas por el mismo parlante, como se puede observar en la figura el pulso glotal tiene cierta similitud entre cada vocal, sin embargo, no son idénticos, ya que como se mencionó anteriormente, la forma del pulso glotal varía ligeramente cuando hay un cambio en el tracto vocal. Es decir, el pulso glotal presenta cierta dependencia respecto al contenido fonético de la voz producida. En la Figura 5.3 se muestran los trenes de pulsos glotales de cuatro parlantes pronunciando la vocal /a/, en la cual se puede apreciar que el

pulso glotal es muy distinto para cada parlante.

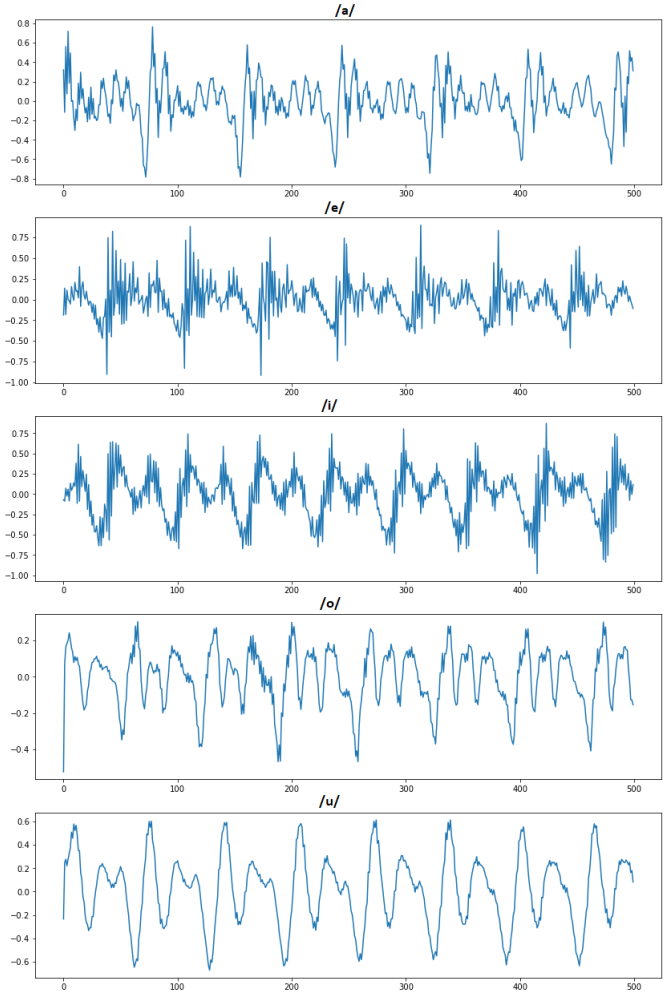

Figura 5.2: Comparación del tren de pulsos glotales de las vocales /a/,/e/,/i/,/o/ y /u/, pronunciadas por el mismo parlante.

El entrenamiento de la CNN es ejecutado bajo un entorno Cloud (Google Colab) que permite mayor potencia al momento de entrenar modelos que demanden gran cantidad de recursos, ya que este en su forma gratuita te permite usar un servidor que cuenta con una unidad de procesamiento gráfico (GPU, por sus siglas en inglés), claro con una cantidad de datos limitada. Los parámetros utilizados para el entrenamiento de la CNN son los siguientes: optimizador *Adam*, como función de pérdida *sparse categorical crossentropy*, para la métrica se usa precisión (*accuracy*), un tamaño de lote (*batch_size*) de 128, para

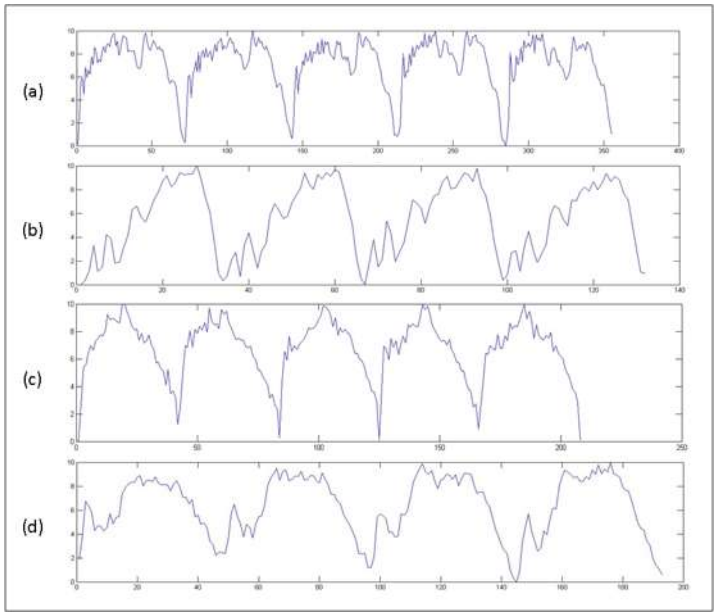

Figura 5.3: Comparación del tren de pulsos glotales de la vocal /a/ pronunciada por distintos parlantes.

combatir el desbalance de los datos se utiliza el ajuste de pesos por clase (*class_weight*), como llamadas de la CNN (*callbacks*) se utilizan: *ModelCheckpoint* que guarda en un archivo el modelo que tuvo un mejor desempeño con los datos de validación durante el entrenamiento. Como decaimiento de la tasa de aprendizaje se utiliza *ReduceLROnPlateau* con un factor de 0.3 y una paciencia de 20 y para la parada temprana (*early stopping*) se utiliza una paciencia de 30.

5.3. Resultados para ELSDSR

Como se menciono en el Capítulo 4, el corpus ELSDSR consta de 22 parlantes (10 mujeres y 12 hombres), donde cada parlante tiene 9 archivos de voz asociados con él/ella. El corpus se divide en dos grupos principales: datos de entrenamiento y datos de prueba, para cada parlante, se usaron 7 archivos de voz para entrenamiento y 2 archivos de voz fueron utilizados para la prueba.

La colección de trenes de pulsos glotales (cada tren de pulsos glotales tiene una longitud de 480 muestras o 30 ms) de los audios de entrenamiento se utiliza para entrenar

el modelo CNN. En la Figura 5.4 se muestra la gráfica de desempeño y de pérdida durante el entrenamiento. En estas figuras se puede ver como gracias a la técnica *parada temprana* se detiene el entrenamiento cuando este ya no mejora, sin embargo se aprecia un poco de sobre-entrenamiento, no obstante al utilizar la función *ModelCheckpoint* (guarda en un archivo el modelo que tuvo un mejor desempeño con los datos de validación durante el entrenamiento) nos aseguramos de quedarnos con el mejor modelo. Las métricas obtenidas en la última época del entrenamiento son las siguientes: para los datos de entrenamiento se obtuvo una *pérdida* de 0.0460 y una *precisión* de 0.9869. Y para los datos de validación se obtuvo una *pérdida* de 0.4323 y una *precisión* de 0.9263 (donde el máximo de *precisión* en este caso es 1.0).

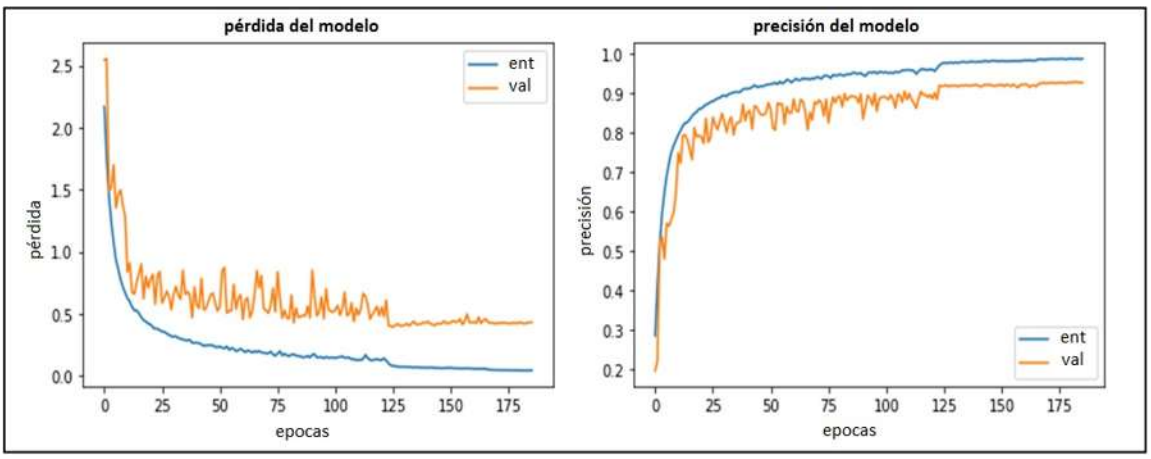

Figura 5.4: Gráficas de desempeño y de costo del modelo CNN durante el entrenamiento del corpus ELSDSR. La gráfica de la izquierda es la función de pérdida y la gráfica de la derecha es la función de desempeño *precisión*.

La *precisión* obtenida para los marcos (480 muestras) de tren de pulsos glotales de los audios de prueba es de 81.58 %. Algo importante a considerar, es que esta medida es sobre los marcos de todos los audios de prueba y no sobre cada audio en concreto. Además se calcula sus curvas ROC para conocer el desempeño de clasificación del modelo, en la Figura 5.5 se muestran las curvas ROC micro-promedio y macro-promedio, donde se puede apreciar un muy buen desempeño del modelo. El AUC micro-promedio es de 98.92 % y el AUC macro-promedio es de 98.90 %.

Como se mencionó antes, existe una consideración extra para la evaluación con los

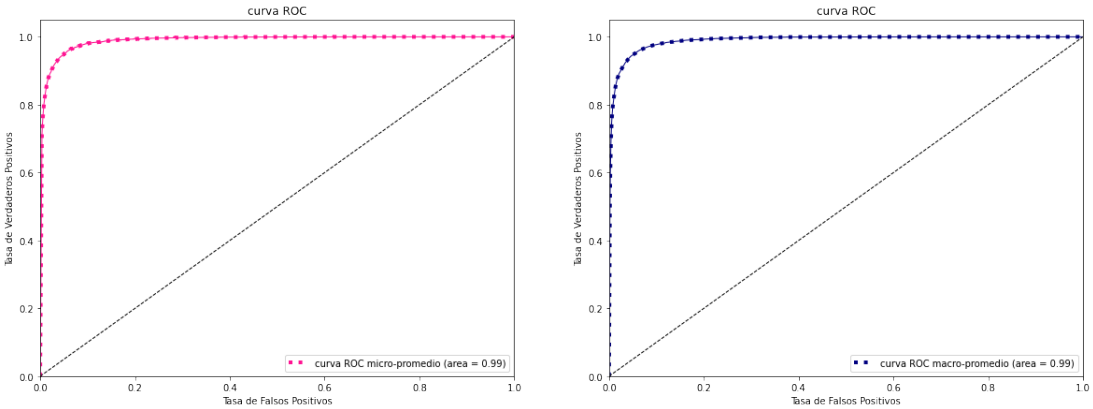

Figura 5.5: Curvas ROC del corpus ELSDSR. La gráfica de la izquierda es la curva ROC micro-promedio y la gráfica de la derecha es la curva ROC macro-promedio.

audios de prueba, ya que la CNN se entrenó con marcos fijos de longitud de 480 elementos (marcos de 30 ms), y los audios no contienen el mismo número de marcos de tren de pulsos glotales, por lo que después de ser clasificado cada marco por la CNN, se implementa un esquema de votación, donde cada marco vota por un parlante y el parlante con más votos gana. Para el esquema de votación se utiliza la técnica "moda estadística", ya que la moda es aquel valor que, dentro de un conjunto de datos, se repite el mayor número de veces, de esta manera se determina a que clase pertenece el audio (ver Figura 4.6). Ahora se pueden realizar nuevas medidas de desempeño del sistema con cada audio de prueba en especifico y poder comparar con otras técnicas de reconocimiento de parlantes, ya que la mayoría de autores realizan las pruebas con los audios de prueba completos. La *precisión* obtenida para los audios de prueba es de 100 %. Además se crea una matriz de confusión, que se muestra en la Figura 5.6.

Es importante realizar una comparativa con las técnicas que han tenido buenos resultados a lo largo de la historia con esta base de datos. En la Tabla 5.1 se muestra la comparación con otros sistemas de reconocimiento de parlantes, en esta tabla se puede apreciar que nuestra técnica supera a las técnicas que han tenido mejores resultados a lo largo de la historia y solo se iguala a la técnica de Rawahy [Al-Rawahy10].

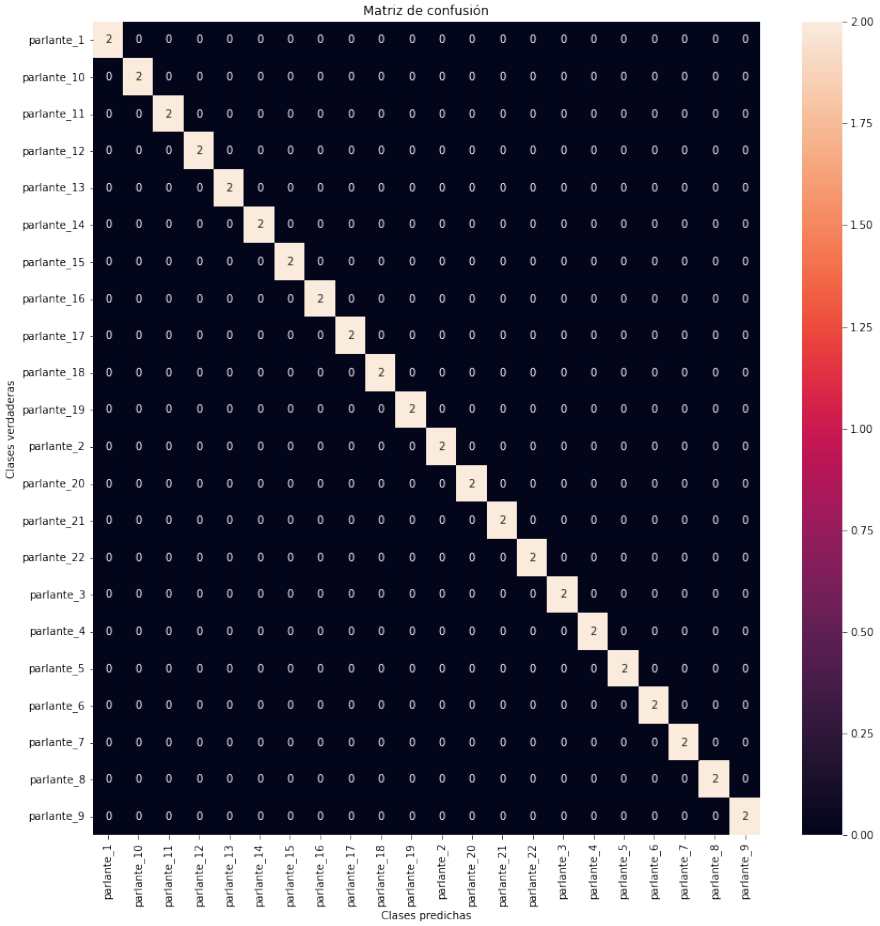

Figura 5.6: Matriz de confusión para el corpus ELSDSR.

Autor	Método	Resultado %
[Castro18]	Formantes	90 %
[Saady14]	WPT	95.7 %
[Al-Rawahy10]	Transformada Zak	100 %
Propuesto	Pulso Glotal	100 %

Tabla 5.1: Comparación con otros métodos (ELSDSR).

5.4. Resultados para TIMIT

Como se mencionó en el Capítulo 4, el corpus TIMIT consta de 630 parlantes (438 hombres y 192 mujeres), donde cada parlante tiene 10 archivos de voz asociados con él/ella. El corpus se divide en dos grupos principales: datos de entrenamiento y datos de prueba,

para cada parlante, se usaron 8 archivos de voz para entrenamiento y 2 archivos de voz fueron utilizados para la prueba.

La colección de trenes de pulsos glotales de los audios de entrenamiento se utiliza para entrenar el modelo CNN. Las métricas obtenidas en la última época del entrenamiento son las siguientes: para los datos de entrenamiento se obtuvo una *pérdida* de 0.9156 y una *precisión* de 0.7481. Y para los datos de validación se obtuvo una *pérdida* de 0.6353 y una *precisión* de 0.8267 (donde el máximo de *precisión* en este caso es 1.0).

La *precisión* obtenida para los marcos (480 muestras) de tren de pulsos glotales de los audios de prueba es de 63.16 %. Algo importante a considerar, es que esta medida es sobre los marcos de todos los audios de prueba y no sobre cada audio en concreto. Además se calcula sus curvas ROC para conocer el desempeño de clasificación del modelo, en la Figura 5.7 se muestran las curvas ROC micro-promedio y macro-promedio, donde se puede apreciar un muy buen desempeño del modelo. El AUC micro-promedio es de 99.53 % y el AUC macro-promedio es de 99.48 %.

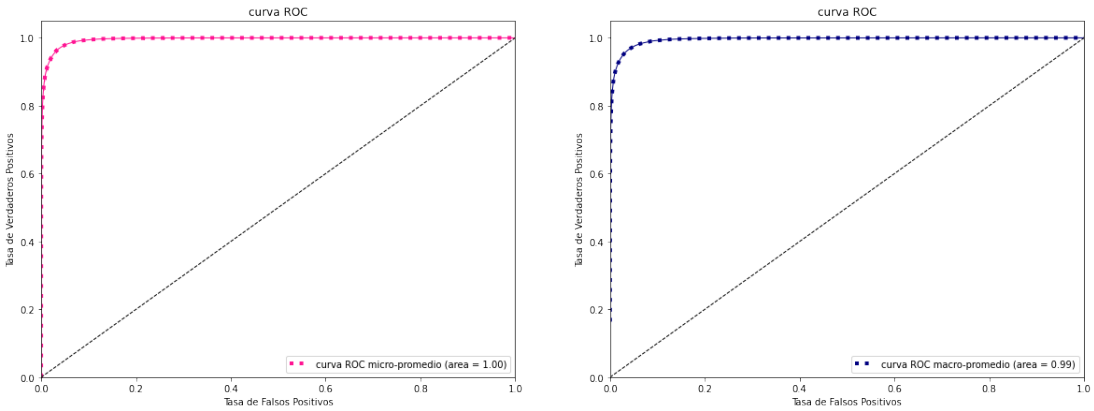

Figura 5.7: Curvas ROC del corpus TIMIT. La gráfica de la izquierda es la curva ROC micro-promedio y la gráfica de la derecha es la curva ROC macro-promedio.

Para los audio de prueba (cada audio de prueba en especifico), se obtuvo una *precisión* de 99.52 %.

En la Tabla 5.2 se muestra la comparación con otros sistemas de reconocimiento de parlantes que han tenido buenos resultados a lo largo de la historia con esta base de datos, en esta tabla se puede apreciar que nuestra técnica supera a las técnicas que han

tenido mejores resultados a lo largo de la historia y solo se iguala a la técnica de Reynolds [Reynolds95].

Autor	Método	Resultado %
[Plumpe99]	Pulso Glotal	70 %
[Thyes00]	PCA-MLES	92.5 %
[Veena15]	Multitaper-MFCC	95 %
[Al-Rawahy12]	DCT-cepstrum	99 %
[Reynolds95]	MFCC-GMM	99.5 %
Propuesto	Pulso Glotal	99.5 %

Tabla 5.2: Comparación con otros métodos (TIMIT).

5.5. Conclusiones del capítulo

Al realizar algunas pruebas, comparando el pulso glotal de un mismo parlante pronunciando distintos sonidos vocalizados y distintos parlantes pronunciando un mismo sonido vocalizado, podemos observar que existen semejanzas para los pulsos glotales de un mismo parlante y diferencias con otros parlantes. Además, de acuerdo a los resultados obtenidos, se confirma que el método propuesto para la estimación del pulso glotal realiza una estimación correcta del pulso glotal y el modelo CNN diseñado una correcta extracción de características y clasificación.

La métrica *precisión* para el caso de datos desbalanceados no resulta ser la mejor elección, sin embargo, al entregar un solo número resulta bastante cómoda para interpretar el desempeño del modelo de una manera rápida. En cambio, la curva ROC y la matriz de confusión nos dan una mejor forma de interpretar el desempeño del modelo, pero a un costo computacional mayor.

Capítulo 6

Conclusiones y Trabajos Futuros

6.1. Conclusiones generales

Gracias a los resultados obtenidos durante las pruebas de nuestro sistema, se puede aseverar que es posible reconocer individuos por medio del pulso glotal de su voz y de CNN 1D, sin importar la frase que estos digan.

Nuestro sistema de reconocimiento de parlantes obtuvo excelentes resultados con las bases de datos ELSDSR y TIMIT, se obtuvo 100 % y 99.52 % de *precisión* respectivamente. Por lo tanto, se puede tener certeza que el proceso que diseñamos para la estimación del pulso glotal y su clasificación por medio de CNN 1D es el correcto. Sin embargo, el pulso glotal presenta cierta dependencia respecto al contenido fonético de la voz producida, por lo tanto, se necesitan que la palabra o frase que se desea identificar contenga distintos sonidos vocalizados, y así tener una mayor fiabilidad de la identificación del parlante.

6.2. Trabajos Futuros

Nuestro sistema de reconocimiento de parlantes por medio del pulso glotal y CNN 1D, contiene varias etapas o partes y variables en su proceso que pueden ser modificados, por lo que el sistema propuesto está disponible a mejoras.

1. Para la detección de los segmentos con sonidos vocalizados se utiliza el método error

de predicción, sin embargo, existen más métodos para su detección y estos pueden ser utilizados.

2. El método propuesto para la estimación del pulso glotal, adicionalmente calcula la señal del tracto vocal durante su proceso y de una manera robusta, esta podría usarse en combinación con el pulso glotal para la identificación de parlantes.

3. Utilizar otra técnica para la estimación del pulso glotal y observar el comportamiento del sistema de reconocimiento.

4. Diseñar y probar el sistema con otro tipo de red neuronal o con una arquitectura del modelo CNN diferente.

5. Utilizar otro tipo de bases de datos como: con un mayor número de parlantes, otro tipo de idioma y con una calidad distinta, por ejemplo, la base de datos NTIMIT que es la base de datos TIMIT con calidad telefónica.

6. Probar el sistema en casos en que se ha modificada la voz del parlante, ya sea por una enfermedad simple como gripa o por ronquera. Del mismo modo, el paso del tiempo y el envejecimiento modifican la voz.

Referencias

[A.02] A., M. C. y Jackson-Menaldi. *La voz patológica.* Ed. Médica Panamericana, 2002.

[Airas08] Airas, M. *Methods and studies of laryngeal voice quality analysis in speech production.* Helsinki University of Technology, Espoo, Finland, 2008.

[Al-Rawahy10] Al-Rawahy, A. A text–independent speaker identification system based on the zak transform. *Signal Processing : An International Journal*, 4, 06 2010.

[Al-Rawahy12] Al-Rawahy, S., Hossen, A., y Heute, U. Text-independent speaker identification system based on the histogram of dct-cepstrum coefficients. *International Journal of Knowledge-based and Intelligent Engineering Systems*, 16, 2012. doi:10.3233/KES-2012-0239.

[Alku92a] Alku, P. *An Automatic method to Estimate the Time-based Parameters of the Glottal Pulseform.* ICASSP, San Francisco, 1992.

[Alku92b] Alku, P. *Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering.* Speech Communication 11, pp. 109-118, 1992.

[Alzamendi16] Alzamendi, G. *Modelado estocástico de la fonación y señales biomédicas relacionadas: Métodos en espacio de estados aplica-*

dos al análisis estructural, al modelado de la fonación y al filtrado inverso. Tesis Doctoral, 06 2016.

[Banerjee18] Banerjee, A., Dubey, A., Menon, A., Nanda, S., y Nandi, G. C. *Speaker Recognition using Deep Belief Networks to CCIS Proceedings.* Robotics and Artificial Intelligence Laborotary, Indian Institute of Information Technology, Allahabad, India, 2018.

[Bozkurt05] Bozkurt, B., Doval, B., d'Alessandro, C., y Dutoit, T. *Zeros of Z-Transform representation with application to source-filter separation in speech.* IEEE Signal Processing Letters, vol. 12, nº 4, pp. 344-347, 2005.

[Brownlee19] Brownlee, J. *Better Deep Learning.* 2019.

[Brunel19] Brunel, A., Pasquet, J., Pasquet, J., Rodriguez, N., Comby, F., Fouchez, D., y Chaumont, M. A cnn adapted to time series for the classification of supernovae, 01 2019.

[Caicedo09] Caicedo, E. y López, J. *Una aproximación práctica a las redes neuronales artificiales.* Santiago de Cali: programa editorial Universidad del valle, 2009.

[Castro18] Castro, M., Camarena, J., y Figueroa, K. Cloud point matching for text-independent speaker identification. *IEEE International Autumn Meeting on Power, Electronics and Computing*, 2018.

[Champod00] Champod, C. y Meuwly, D. The inference of identity in forensic speaker recognition. *Speech Communication*, 31(2):193–203, 2000. ISSN 0167-6393. doi: https://doi.org/10.1016/S0167-6393(99)00078-3.
URL `https://www.sciencedirect.com/science/article/pii/S0167639399000783`

[Chi07] Chi, X. y Sonderegger, M. *Subglottal coupling and its influence on vowel formants.* The Journal of the Acoustical Society of America, 2007.

[Cobeta13] Cobeta, I., Núñez, F., y Fernández, S. *Patología de la voz.* ICG Marge, SL, 2013.

[Degottex10] Degottex, G. *Glottal source and vocal-tract separation. Signal and Image processing.* Université Pierre et Marie Curie - Paris VI, 2010.

[denBerg58] den Berg, J. V. *Myoelastic-aerodynamic theory of voice production.* J Speech Hear Res, 1958.

[deOliveira Dias12] de Oliveira Dias, S. *Estimation of the glottal pulse from speech or singing voice.* Proyecto Fin de Carrera, 2012.

[Dias12] Dias, S. y Ferreira, A. *GLOTTAL PULSE ESTIMATION - A FREQUENCY DOMAIN APPROACH.* Department of Electrical and Computer Engineering, University of Porto - Faculty of Engineering, Porto, Portugal, 2012.

[Drugman09] Drugman, T., Bozkurt, B., y Dutoit, T. *Complex Cepstrum-based Decomposition of Speech for Glottal Source Estimation.* Interspeech, pp. 116- 119, 2009.

[Duxans00] Duxans, H. y Bonafonte, A. *Revisión de técnicas de estimación de pulso glotal basadas en filtrado inverso.* Universitat Politècnica de Catalunya, Barcelona, 2000.

[Fant60] Fant, G. *Acoustic Theory of Speech Production.* Mouton, The Hague, 1960.

[Fant79] Fant, G. *Vocal-source analysis.* a progress report. STL-QPSR, 20 (3-4): pp. 31-53, 1979.

[Fant85] Fant, G., Liljencrants, J., y guaq Lin, Q. *Vocal-source analysis. A four parameter model of glottal flow.* STL-QPSR, 4, pp. 1-13, 1985.

[Fant93] Fant, G. *Some problems in voice source analysis.* Speech Communication, 1993.

[Fant95] Fant, G. *The LF-model revisited. Transformations and frequency analysis.* STL-QPSR, 36 (2-3): pp. 119-156, 1995.

[Feng05] Feng, L. y Hansen, L. K. A new database for speaker recognition. Inf. téc., Informatics and Mathematical Modeling, Technical University of Denmark, 2005.

[Flanagan72] Flanagan, J. L. *Speech Analysis, Synthesis and Perception.* 2nd ed., Springer, 1972.

[Goodfellow16] Goodfellow, I. J., Bengio, Y., y Courville, A. *Deep Learning.* MIT Press, Cambridge, MA, USA, 2016.

[Guitart04] Guitart, J. M. *Sonido y sentido.* Georgetown University Press, 2004.

[Guonason08] Guonason, J. y Brookes, M. Voice source cepstrum coefficients for speaker identification. págs. 4821 – 4824. 05 2008. doi:10.1109/ICASSP.2008.4518736.

[Géron17] Géron, A. *Hands-On Machine Learning with Scikit-Learn and TensorFlow.* 2017.

[Hagan14] Hagan, M., Demuth, H., Beale, M., y Jesús, O. D. *Neural Network Design.* 2014.

[Home22] Home, N. *Oraciones para base de datos ELSDSR.* Accedido por ultima vez el 05/07/2022.
URL `http://www.pbs.org/wgbh/nova/pyramid/`

[Ibarrola11] Ibarrola, J. A. C. Notas de síntesis y reconocimiento de voz. Inf. téc., Universidad Michoacana de San Nicolás de Hidalgo, 2011.

[Ioffe15] Ioffe, S. y Szegedy, C. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.* 2015.

[Javkin87] Javkin, H. R., Antõnanzas-Barroso, N., y Maddieson, I. *Digital Inverse Filtering for Linguistic Research.* Journal of Speech and Hearing Research 30, pp. 122-129, 1987.

[kafentzis08] kafentzis, G. P. *On the glottal flow derivative wavefom and its properties.* Bachelor's Dissertation, University of Crete, Greece, 2008.

[Kingma15] Kingma, D. P. y Ba, J. L. *ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION.* 2015.

[Lao León17] Lao León, Y., Rivas Méndez, A., Pérez Pravia, M., y Marrero Delgado, F. Procedimiento para el pronóstico de la demanda mediante redes neuronales artificiales / procedure for forecasting demand by using artificial neural networks. *Ciencias Holguin*, 23:43–59, 01 2017.

[Lin14] Lin, M., Chen, Q., y Yan, S. *Network In Network.* 2014.

[Miyara22a] Miyara, F. La naturaleza del sonido. Inf. téc., Universidad Nacional de Rosario, Accedido por ultima vez 05/05/2022.
URL `https://www.fceia.unr.edu.ar/acustica/biblio/sonido.htm`

[Miyara22b] Miyara, F. La voz humana. Inf. téc., Universidad Nacional de Rosario, Accedido por ultima vez 05/05/2022.
URL `https://www.fceia.unr.edu.ar/acustica/biblio/fonatori.pdf`

[Murphy08] Murphy, K. *Digital signal processing techniques for application in the analysis of pathological voice and normophonic singing voice.* Tesis Doctoral, 2008.

[Nakagawa12] Nakagawa, S., Wang, L., y Ohtsuka, S. Speaker identification and verification by combining mfcc and phase information. *IEEE Transactions on Audio, Speech Language Processing*, 20:1085–1095, 05 2012. doi:10.1109/TASL.2011.2172422.

[Obediente98] Obediente, E. *Fonética y fonología.* Universidad Los Andes, 1998.

[Perelló62] Perelló, J. *La théorie muco-ondulatoire de la phonation.* Ann NY Acad Sci, 1962.

[Plumpe99] Plumpe, M. D., Quatieri, T. F., y Reynolds, D. A. *Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification.* IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 7, NO. 5, 1999.

[Rabiner11] Rabiner, L. R. y Schafer, R. W. *Digital Speech Processing.* Prentice Hall, 2011.

[Reynolds95] Reynolds, D. A. *Large Population Speaker Identification Using Clean and Telephone Speech.* IEEE SIGNAL PROCESSING LETTERS, VOL. 2, NO. 3, 1995.

[Reynoso21] Reynoso, M. *Identificación de Parlantes Independiente del Texto Utilizando Redes Neuronales Convolucionales.* Proyecto Fin de Carrera, 2021.

[Rosenberg71] Rosenberg, A. E. *Effect of glottal pulse shape on the quality of natural vowels.* Journal of the Acoustical Society of America, 49(2B):583–590, 1971.

[Saady14] Saady, M., El-Borey, H., El-Dahshan, E.-S., y Yahia, a. s. Standalone intelligent voice recognition system. *Journal of Signal and*

Information Processing, 05:179–190, 01 2014. doi:10.4236/jsip.2014.54019.

[Saks97] Saks, M. J. *Merlin and solomon: Lessons from the law's formative encounters with forensic identification science. Hastings Lj, 49:1069.* 1997.

[scikitlearndevelopers22] scikit-learn developers. *Receiver Operating Characteristic (ROC).* Accedido por ultima vez el 05/07/2022.
URL `https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html`

[Shannon49] Shannon, C. *Communication in the presence of noise.* Proceedings of the Institute of Radio Engineers, 1949.

[Shenfield20] Shenfield, A. y Howarth, M. A novel deep learning model for the detection and identification of rolling element-bearing faults. *Sensors (Basel, Switzerland)*, 20, 09 2020. doi:10.3390/s20185112.

[Soong85] Soong, F., Rosenberg, A., Rabiner, L., y Juang, B.-H. Report: A vector quantization approach to speaker recognition. tomo 66, págs. 387 – 390. 05 1985. doi:10.1109/ICASSP.1985.1168412.

[Stevens98] Stevens, K. N. *Acoustic Phonetics.* MIT Press, Cambridge, MA, 1998.

[Sundberg87] Sundberg, J. *The science of singing voice.* Northern Illinois, University Press. Dekalb, Illinois, 1987.

[Thyes00] Thyes, O., Nguyen, P., y Junqua, J.-C. Speaker identification and verification using eigenvoices. págs. 242–245. 01 2000.

[TI22] TI, SRI, y MIT. *TIMIT.* Accedido por ultima vez el 05/07/2022.
URL `https://catalog.ldc.upenn.edu/LDC93s1`

[Titze08] Titze, I. R. *Nonlinear source–filter coupling in phonation: Theory.* The Journal of the Acoustical Society of America, 2008.

[Torres07] Torres, B. *Anatomía funcional de la voz*, págs. 1–21. 01 2007.

[Torres18] Torres, J. *Deep Learning, Introducción práctica con Keras (PRIMERA PARTE).* Universidad Politécnica de Catalunya - UPC Barcelona Tech, 2018.

[Veena15] Veena, K. y Mathew, D. Speaker identification and verification of noisy speech using multitaper mfcc and gaussian mixture models. *2015 International Conference on Power, Instrumentation, Control and Computing (PICC)*, págs. 1–4, 2015.

[Veeneman85] Veeneman, D. y BeMent, S. *Automatic Glottal Inverse Filtering from Speech and Electroglottographic Signals.* IEEE Trans. Acoustics, Speech and Signal Processing, vol. ASSP-33, 1985.

[Zhang21] Zhang, B., Shi, Y., Hou, L., Yin, Z., y Chai, C. Tsmg: A deep learning framework for recognizing human learning style using eeg signals. *Brain Sciences*, 11:1397, 10 2021. doi: 10.3390/brainsci11111397.