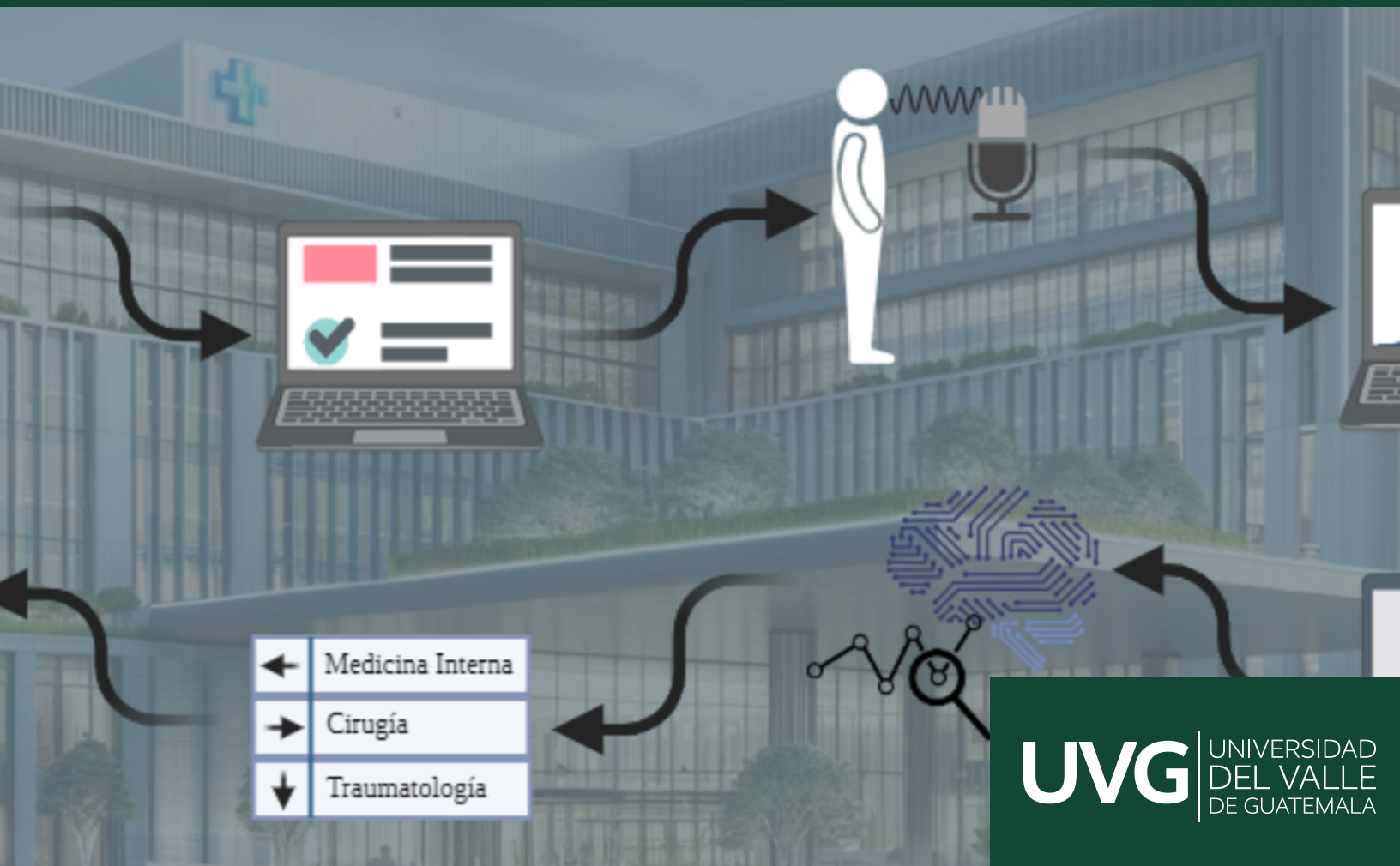
Desarrollo de algoritmo para la realización de entrevistas médicas preliminares por reconocimiento y transcripción de voz

Andrés Alejandro Mérida Ruano



UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería

Desarrollo de algoritmo para la realización de entrevistas médicas preliminares por reconocimiento y transcripción de voz

Trabajo de graduación presentado por Andrés Alejandro Mérida Ruano
para optar al grado académico de Licenciado en Ingeniería Biomédica

Guatemala,

2025

UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería

Desarrollo de algoritmo para la realización de entrevistas médicas preliminares por reconocimiento y transcripción de voz

Trabajo de graduación presentado por Andrés Alejandro Mérida Ruano
para optar al grado académico de Licenciado en Ingeniería Biomédica

Guatemala,

2025

Vo.Bo.:

(f) _____

M. Sc. Carlos Esquit

Tribunal Examinador:

(f) _____

M.Sc. Carlos Esquit

(f) _____

Dr. Luis Alberto Rivera Estrada

(f) _____

Ing. Kurt Emmanuel Kellner

Fecha de aprobación: Guatemala, 13 de febrero de 2025.

Prefacio

Este proyecto nace de la visión de poder utilizar todos los conocimientos y la tecnología disponible en la actualidad para poder, no solo solventar problemáticas actuales del sistema de salud, sino también para facilitar el trabajo de las personas de este sector. Nosotros, como ingenieros biomédicos, hemos tenido la posibilidad de ver de cerca la complejidad que tiene llevar a cabo el trabajo de todas las personas que conforman esta área. Es de conocimiento general que el sistema de salud guatemalteco tiene una gran cantidad de carencias y áreas de mejora y nuestra labor es con nuestras habilidades poder cambiar esta situación de la mejor manera que podamos. Al final del día una gran cantidad de nosotros, incluyéndome, entró a esta carrera porque siente que tiene algo que aportar en la medicina desde la ingeniería y este es únicamente el comienzo de ese gran camino que cada uno de nosotros quiere emprender.

La idea de un algoritmo para realización de entrevistas médicas preliminares por reconocimiento y transcripción de voz surgió gracias a una conversación con un amigo perteneciente al sector de salud. Él menciona que en una gran cantidad de veces la clasificación de pacientes en los hospitales es llevada a cabo por personas que no están precisamente capacitadas para esto. Esto derivaba en que se generan muchos errores, esto no suele ser un problema muy grave, ya que luego se pueden re clasificar, pero la misma falta de personal produce que se tenga que recurrir a esto. De igual forma, vuelve el sistema aún más ineficiente y una de las formas en las cuales esto puede ser solucionado es que este proceso sea llevado a cabo de manera computacional. La parte de reconocimiento de voz surge debido a que no todas las personas tienen la capacidad de utilizar una computadora para escribir las respuestas a la entrevista. Es mucho más fácil si se expresan de manera oral.

Me gustaría agradecer primero a mis padres que me han apoyado en todas las formas posibles para que hoy me encuentre presentando mi trabajo de graduación y esté cerrando una carrera en la que he aprendido infinidad de cosas. También quiero agradecer a mi hermana que ha estado para mí en este último año dándome fuerzas para seguir. A mi familia en general, que está muy orgullosa de todo lo que se ha conseguido durante este tiempo y con este proyecto en específico. Quiero agradecer a Dulce, Renata, Rodolfo, Silvana y Sofía que me han acompañado desde hace mucho, con quienes hemos sufrido juntos y nos hemos apoyado cada vez que lo necesitábamos, que hemos crecido juntos tanto a nivel personal como profesional y han sido mi apoyo para la construcción de este proyecto.

También me gustaría agradecer a todos los que formaron parte de las pruebas llevadas a cabo a lo largo de la experimentación de este proyecto, sin los cuales no se habrían obtenido los resultados que hoy presentamos. Quiero agradecerle al ingeniero Luis Alberto Rivera por su ayuda en la última parte de este proyecto, a todas las personas del sector de salud que fueron encuestados. A mi amigo estudiante de medicina que contribuyó y me ayudó a poder entender de una mejor forma toda la parte de medicina involucrada. Agradecer al departamento y a la universidad por todos estos años de aprendizaje, de acompañamiento, de experiencias y de crecimiento que me han brindado. Por último, agradecer a Dios por darme esta gran oportunidad en la vida.

Índice

Lista de figuras

Resumen

El sistema de salud de Guatemala, principalmente en el sector público, cuenta con una densidad de personal médico por debajo de los estándares mínimos necesarios para operar de manera adecuada. Esto produce, no solo insatisfacción general en los pacientes que acuden a los centros de salud, sino también una alta carga de trabajo sobre todos los trabajadores que forman parte del personal médico. Esta alta carga genera una mayor tasa de incidentes peligrosos para la salud en el área de trabajo y una menor calidad en el servicio prestado a los pacientes. Un tiempo alto de espera previo a una consulta se encuentra ligado al uso de los recursos dentro del hospital, que en este caso son reducidos y esto tiene implicaciones negativas sobre el sistema completo incluyendo a los pacientes que acceden de forma tardía a sus tratamientos.

Se desarrollaron diferentes versiones de algoritmos de reconocimiento y transcripción de voz y se compararon con base en el porcentaje de efectividad de transcripción bajo diferentes condiciones. Se llevaron a cabo encuestas a personas del personal médico para formar el esquema de la entrevista. Por último, se desarrolló un algoritmo de clasificación que utilizaba las respuestas de la entrevista para determinar a qué área debía asignarse el paciente. Se logró desarrollar un algoritmo de reconocimiento de voz que presentó una efectividad casi ideal. El modelo de clasificación tiene precisión alta al evaluarla con casos de entrenamiento y limitada con casos reales. La versión final es capaz de generar un reporte final con la entrevista transcrita y la clasificación sugerida.

Abstract

The health system in Guatemala, mainly in the public sector, has a density of medical personnel below the minimum standards necessary to operate adequately. This produces not only general dissatisfaction among patients attending health centers, but also a high workload on all workers who are part of the medical staff. This high workload generates a higher rate of health hazardous incidents in the work area and a lower quality of service provided to patients. A high waiting time prior to a consultation is linked to the use of resources within the hospital, which in this case are reduced and this causes negative implications on the entire system including patients who access their treatments late.

Different versions of speech recognition and transcription algorithms were developed and compared based on the percentage of transcription effectiveness under different conditions. Surveys of medical personnel were conducted to form the interview scheme. Finally, a classification algorithm was developed that used the interview responses to determine to which area the patient should be assigned. We were able to develop a speech recognition algorithm that had near-ideal effectiveness. The classification model has high accuracy when evaluated with training cases and limited accuracy with real cases. The final version is able to generate a final report with the transcribed interview and the suggested classification.

CAPÍTULO 1

Introducción

La disponibilidad de personal médico en los servicios de salud utilizables por la población es un problema que debe de ser abordado ya que es uno de los aspectos más importantes en la vida de las personas. Esta misma disponibilidad tiene efectos sobre la calidad de servicio que se le pude brindar a un paciente. En Guatemala, la cantidad de trabajadores en sanidad para cubrir las necesidades de la población no es la adecuada lo cual imposibilita llevar muchos procesos básicos de manera eficiente. Cuando un médico o enfermero se encuentra limitado en el tiempo que le puede dedicar a una entrevista preliminar se abre la puerta a errores en lo que puede ser su clasificación. Contar con la historia completa de la perspectiva del paciente es algo que puede desaparecer cuando alguien ajeno al paciente redacta esta información transmitida. Toda la información brindada por el paciente es importante y muchas veces dentro de las entrevistas el entrevistados podría llegar a omitir algo de información por el simple error humano.

El proyecto se dividió en tres partes fundamentales. El esquema de la entrevista, el algoritmo de reconocimiento y la transcripción de voz y el algoritmo para realizar la entrevista y generar un reporte con la especialidad a visitar. La primera se trabajó realizando encuestas a personal médico para determinar como organizarían una entrevista. A raíz de esto se determinó el esquema de la entrevista de 10 preguntas en orden adecuado y con base en cuáles fueron consideradas de mayor relevancia. Para el primer algoritmo, se llevaron a cabo diferentes pruebas con sujetos para media la efectividad de transcripción de diferentes versiones de algoritmos. Al contar con el algoritmo final se realizaron pruebas para medir diferentes parámetros de su adaptabilidad a otros escenarios y su velocidad de transcripción. Por último, se integraron ambas partes en un mismo algoritmo que, con base en las respuestas, generará un reporte final con todas las preguntas, sus respuestas y el especialista al que se recomienda sea asignado el paciente.

En este proyecto, se encuentran diferentes secciones que explican aspectos como la importancia del proyecto, las metas a alcanzar, todos los procesos relevantes en su ejecución y lo conseguido. Comienza por los antecedentes más importantes en el área de reconocimiento de voz y la telemedicina en la actualidad. Sigue la problemática y su respectiva justificación para la realización de este proyecto. Luego, se encuentran los objetivos tanto general como específicos y el alcance que se tiene con este proyecto. Continúa con el marco teórico, donde hay información sobre fundamentos fisiológicos de la voz, el proceso y funcionamiento del reconocimiento de voz, el procesamiento del lenguaje y las entrevistas médicas. Sigue con la metodología empleada en el proyecto y los resultados de toda la experimentación implementada. En las últimas secciones se encuentra la discusión de estos resultados junto a las conclusiones y recomendaciones.

CAPÍTULO 2

Antecedentes

En la actualidad, el reconocimiento de voz ha avanzado tanto que no se utiliza únicamente en los dispositivos de uso diario, sino también en aparatos de diagnóstico. Los dispositivos de cuidado de la salud activados por voz tienen presencia en cuidado remoto, rehabilitación, enfermedades crónicas y otros aspectos relacionados [1]. Uno de los ámbitos con mayor presencia y avances en este tipo de algoritmos de reconocimiento y transcripción en el sector médico es el de radiología. Es una forma en la que se logra reducir el tiempo que transcurre entre el escaneo radiológico y el reporte [2]. Los métodos de estos sistemas aún presentan diferentes limitaciones que afectan la calidad de reconocimiento de cada sistema. Algunos ejemplos de esto son los acentos y velocidad de pronunciación de cada persona, la capacidad de reconocer y diferenciar palabras compuestas, el nivel de entrenamiento y volumen del diccionario que tiene para el reconocimiento y la precisión para realizar la transcripción en tiempo real. Todas estos problemas afectan en distinta medida con base en la aplicación que se le busca dar [3]. Un posible ámbito de uso donde el reconocimiento de voz puede abrir muchas puertas es la telemedicina.

El crecimiento en la telemedicina ha venido acompañado por la inclusión del reconocimiento de voz en este tipo de tecnología. Uno de los ejemplos más complejos es el de diagnóstico de enfermedades como afonía, insuficiencia cardíaca aguda o enfermedad pulmonar obstructiva por identificación de irregularidades en el habla [4]. El diagnóstico inteligente por *machine learning* y redes neuronales ha sido incluido en softwares y aplicaciones de teléfono para permitir el auto diagnóstico en telemedicina [5]. En Francia se realizó un estudio en estudiantes para conocer su aceptabilidad. Los estudiantes consideraban que puede llegar a ser una tecnología de fácil uso, que, si se realiza de manera correcta, reduce todo a una buena comunicación por medio de instrucciones entre el médico y el paciente. Recalcaron que parte fundamental para que este tipo de tecnología funcione es que exista competencia de desarrollo para evitar su estancamiento [6].

Cuando se busca realizar un prediagnóstico las variables de análisis se encuentran en el texto recopilado del historial clínico y no en imágenes a procesar. Ejemplo de esto es PhenIX (Institute for Medical Genetics and Human Genetics, Berlín, Alemania) que, no utilizando un historial, pero si el texto de la secuencia genética de un paciente es capaz de identificar las variantes de la información, priorizar aquellas que puedan estar directamente ligadas a factores de patogenicidad y con base en esto diagnosticar enfermedades mendelianas. Existe también una herramienta conocida como Xrare (GenomCan Inc., Chengdu, China) que utiliza algoritmos de puntuación por similitud, la información genética almacenada en base de datos y además es capaz de realizar una priorización de las variantes a analizar. Las metodologías utilizadas por estas herramientas poseen una mayor complejidad e incorporan otras variables que complementan el análisis, pero tienen su base en análisis de cadenas de texto y comparación para realizar un diagnóstico [7].

CAPÍTULO 3

Justificación

Guatemala cuenta con un serio problema de disponibilidad de personal médico en relación a la densidad poblacional que requiere de este servicio. En muchas ocasiones, los médicos y enfermeros no se dan abasto para cumplir de manera eficiente con las distintas emergencias o complicaciones que puedan presentar los pacientes. A nivel centroamericano, Guatemala presenta el menor número de trabajadores sanitarios por cada diez mil habitantes, únicamente 12.5 trabajadores. No solo esto, sino que tampoco cumple con los estándares mínimos recomendados por la Organización Internacional de Trabajo para mejorar el beneficio producido sobre la población general y es la mitad de la densidad esperada por la Organización Mundial de la Salud para manejar de manera adecuada el sistema de salud de una nación [8].

La falta de personal médico produce un aumento en los tiempos de espera y hace que las entrevistas se lleven a cabo de manera apresurada lo cual da cabida al error. Uno de los mayores limitantes a la hora de realizar la entrevista inicial con los pacientes es el tiempo con el que cuenta el médico para realizarla. Los pacientes muchas veces no son conscientes del tiempo reducido con el que cuentan los doctores debido a la gran cantidad de otras actividades que deben de realizar. En muchas ocasiones los pacientes buscan acaparar la atención de los médicos y desvían la conversación a temas que no contribuyen a poder generar un diagnóstico correcto [9]. A la hora de realizar una entrevista inicial es importante reconocer que tanto la perspectiva del personal médico que la realiza como la perspectiva del paciente tienen relevancia ya que ambas se complementan para contar la historia completa. La omisión de la perspectiva médica excluye información médica crítica necesaria para el diagnóstico, mientras que la omisión de la perspectiva del paciente suele derivar en la supresión de datos relacionados a su persona que ayudan a la determinación del mejor método terapéutico [10].

Varios estudios han demostrado que los médicos tienden a presentar sesgos de confirmación a la hora de realizar los diagnósticos iniciales de sus pacientes. Este sesgo de confirmación representa la tendencia de una persona a dejarse guiar por la información que confirma su diagnóstico o teoría inicial de una situación, no permitiéndole analizar otras opciones. Este problema en muchos casos termina produciendo diagnósticos erróneos que perjudican al paciente con un tratamiento equivocado por un periodo de tiempo indefinido [11] [12]. El tiempo que transcurre entre la primera visita y el diagnóstico juega un rol muy importante en el tratamiento de la enfermedad del paciente y su debida recuperación. Existen muchos factores que influyen en esto como contar con un médico al que se acude regularmente, un diagnóstico previo diferente, el retraso en referir a un especialista o el acceso a un centro de salud apto. Sin embargo, lo más importante siempre va a ser la comunicación paciente-médico sobre sus síntomas y la capacidad de este último de distinguir aquellos datos y síntomas con mayor relevancia [13].

En este proyecto se propone utilizar algoritmos de reconocimiento y análisis del habla para prevenir este tipo de sesgos, facilitar el acceso a salud a más personas y hacer las entrevistas iniciales a pacientes más eficientes. Con este método, se asegura mantener el modelo oral de entrevista sin excluir a aquellas personas que no manejen el uso de la tecnología o cuenten con alguna imposibilitación de movilidad. De igual manera, este tipo de tecnología facilita este tipo de entrevistas a pacientes que han sufrido pérdida de audición o visión [14]. Para un audio de alta calidad, el acierto de una transcripción automática en comparación con una manual es mayor al 90 %, permitiendo así transformar la forma en la que se realizan las entrevistas actualmente a entrevistas manejadas por una computadora que obtiene toda la información y ya transcrita la transmite al médico para continuar el proceso [15]. En el ámbito médico cualquier detalle de la información brindada por el paciente puede llegar a ser determinante por lo que contar no solo con el reporte de lo dicho sino también con la grabación permite acudir a ella en caso de que sea necesario. La omisión de información por error humano puede suceder y reducir este tipo de problemas es algo que se busca con este proyecto [16].

También se busca agilizar el proceso de la obtención de datos iniciales por medio de estas tecnologías, ya que el tiempo de espera está directamente relacionado con cómo se utilizan los recursos disponibles dentro del centro de salud. Ha quedado demostrado que el aumento en los tiempos de espera trae consecuencias negativas como peores resultados clínicos, insatisfacción y ansiedad por parte de los pacientes, y un acceso tardío a tratamientos. Todas estas repercusiones tienen un efecto directo sobre el bienestar general de la sociedad, lo que a largo plazo afecta de nuevo al sistema de salud [17]. No solo es importante el tiempo que se ahorra dentro de la visita del paciente, sino también la carga de trabajo que se reduce del personal médico, permitiendo así centrarse en la parte más importante del diagnóstico y tratamiento. La alta carga a la que se encuentran sometidos constantemente los doctores tiene efectos adversos en el sistema de salud, el cuidado de los pacientes y la salud del propio doctor [18]. El agotamiento laboral duplica la cantidad de incidentes peligrosos para la salud del paciente, disminuye la calidad de atención prestada y la cantidad de horas trabajadas, abandonando la práctica con una mayor frecuencia que aquellos que no sufren de agotamiento [19].

CAPÍTULO 4

Objetivos

4.1. Objetivo general

Desarrollar un algoritmo para analizar entrevistas preliminares realizadas a pacientes, identificar sus síntomas y sugerir un especialista adecuado para una consulta posterior.

4.2. Objetivos específicos

- Establecer el esquema y preguntas de las entrevistas preliminares a usuarios, realizando encuestas y entrevistas con al menos 5 distintos médicos internistas y especialistas.
- Evaluar la influencia de al menos 3 métodos de grabación en la calidad del audio de las entrevistas.
- Desarrollar un algoritmo para el reconocimiento y transcripción de voz de las entrevistas de usuarios, utilizando librerías y diccionarios de fuente abierta en español.
- Integrar el esquema y reconocimiento de voz en un algoritmo capaz de analizar entrevistas a usuarios de prueba y generar un reporte final a partir de las respuestas, incluyendo los síntomas detectados y los posibles especialistas a visitar.

CAPÍTULO 5

Alcance

El proyecto tiene como objetivo principal el desarrollo de un algoritmo que sea capaz de llevar a cabo una entrevista médica preliminar y obtenga las respuestas por medio de reconocimiento de voz. Se enfocará en que los síntomas sean obtenidos dentro de las preguntas realizadas y poder generar una sugerencia de cuál debería ser el área a la cual debe avocarse el paciente con base en su entrevista. El esquema de las preguntas se obtendrá por medio de encuestas realizadas a una muestra pequeña de personas pertenecientes al sector de salud y se tomará como parámetro principal la recurrencia con la que las preguntas aparecen en los esquemas escogidos por estos profesionales. Este esquema únicamente contará con diez preguntas, por lo que no se puede abarcar todo el espectro de posibles consultas, sino que se buscan los puntos más importantes a tocar durante una entrevista para que conformen la estructura de preguntas.

Los objetivos relacionados a reconocimiento y transcripción de voz tendrán como idea principal determinar a base de pruebas el método más eficaz para realizar este proceso. Se evalúan diferentes versiones de algoritmos variando la forma de adquisición de audio, procesamiento de la señal o sistema de transcripción para concluir cual de todos es el más adecuado para ser implementado dentro del algoritmo final. Las pruebas que se llevan a cabo para realizar esta distinción son de efectividad de transcripción únicamente. Luego, con la versión elegida, se miden otros parámetros relevantes como la velocidad de transcripción y la diferencia de efectividad entre un reconocimiento de voz con una lectura textual o la forma regular del habla.

En el algoritmo final se incluyen el esquema de la entrevista, el reconocimiento de voz y un sistema de predicción para generar una sugerencia de cuál podría ser el área a la cual asignar al paciente en cuestión. Todo esto servirá para generar un reporte final con todas las preguntas, sus respuestas, incluyendo dentro de ellas los síntomas reportados por el paciente y la sugerencia basada en las respuestas dadas. Este proyecto únicamente estará enfocado en entrevistas realizadas en idioma español, la cantidad de personas involucradas en las encuestas como en los experimentos relacionados a la verificación del funcionamiento de los distintos algoritmos será limitada. No se tomarán en cuenta las diferencias que existen en

acentos o dialectos presentes dentro de esta población. No se utilizarán sistemas complejos de inteligencia artificial más allá de los utilizados en el análisis básico de respuestas y el reconocimiento de voz. La cantidad de muestras utilizadas para el entrenamiento y prueba del algoritmo para las sugerencias de especialidades será limitada. Todo lo obtenido de las pruebas será utilizado bajo la aprobación de los individuos involucrados y se espera que sirva como base para futuras aplicaciones más complejas en la rama.

CAPÍTULO 6

Marco teórico

6.1. La voz

La voz se genera gracias al proceso de fonación que se produce en la laringe y que a partir de la vibración de las cuerdas vocales es capaz de externar el sonido desde el cuerpo hacia nuestro entorno. Este fenómeno es utilizado por los humanos para expresarse y comunicarse de manera verbal con los demás. Existen otros fenómenos que la complementan como la modulación y la proyección del aire al momento de la espiración. La producción de la voz cuenta con diferentes procesos dentro de los cuales los más importantes son la respiración, la fonación, la resonancia y la articulación [20].

6.1.1. Fundamentos fisiológicos

La respiración es el paso del aire que es espirado por los pulmones y producido gracias a la contracción de los músculos. Esta contracción genera que el aire dentro de los pulmones sea expulsado y utilizarlo para los siguientes procesos en la producción de voz. La fonación tiene lugar en la laringe, más específicamente en las cuerdas vocales. Estas vibran gracias al aire espirado que atraviesa la zona y estas vibraciones son las que en forma de ondas generan el sonido primario. Características como la calidad y tono vienen dadas por la tensión, longitud y grosor de estas. El fenómeno físico de la resonancia sucede en las cavidades de la faringe, oral y nasal, aquí el sonido se modifica y amplifica transformándolo en lo que ya corresponde al sonido de la voz. Por último, la parte más importante relacionada a la transmisión es la articulación ya que los labios, dientes, lengua y paladar modulan la voz para pronunciar los diferentes fonemas que constituyen el lenguaje con el que nos comunicamos [20].

Uno de los principales reguladores en la generación de la voz es la presión subglótica, esta se encarga de controlar distintos parámetros involucrados y es un iniciador en el proceso de vibración de las cuerdas vocales. Esta presión es generada a partir de la regulación de presión alveolar de la cual un porcentaje se ubica justo por debajo de las cuerdas vocales y

así consigue su movilización utilizando el flujo de aire. El volumen pulmonar juega un papel muy importante en la producción de voz ya que se requieren diferentes niveles de fuerza activa y pasiva de los músculos de la cavidad torácica dependiendo del volumen presente. Por ejemplo, para que el tono de voz sea el adecuado y cómodo para la persona se necesita un volumen de aproximadamente 4-5 cm de agua. Es importante tomar en cuenta que el aumento del tono representa también un aumento de la presión subglótica por lo que variar el tono significa trabajo en los músculos abdominales para producir una inhalación activa con la cual conseguir ese volumen requerido [21].

El aparato encargado de producir la voz y el habla es el aparato fonador (Figura 1). Los órganos presentes en este proceso de creación se pueden dividir en tres subgrupos principales: las cavidades infraglóticas que son todos los órganos de respiración como los pulmones, tráquea, bronquios, músculos intercostales y diafragma; la cavidad laríngea donde se encuentran las cuerdas vocales, uno de los principales órganos de producción de la voz y la epiglotis. Al final de todo este conducto encontramos las cavidades supraglóticas que están formadas por la cavidad bucal, nasal y la faringe. La faringe une a la laringe con las demás cavidades y estas acaban en los labios que realizan movimientos de apertura o cierre ayudando a la articulación [22].

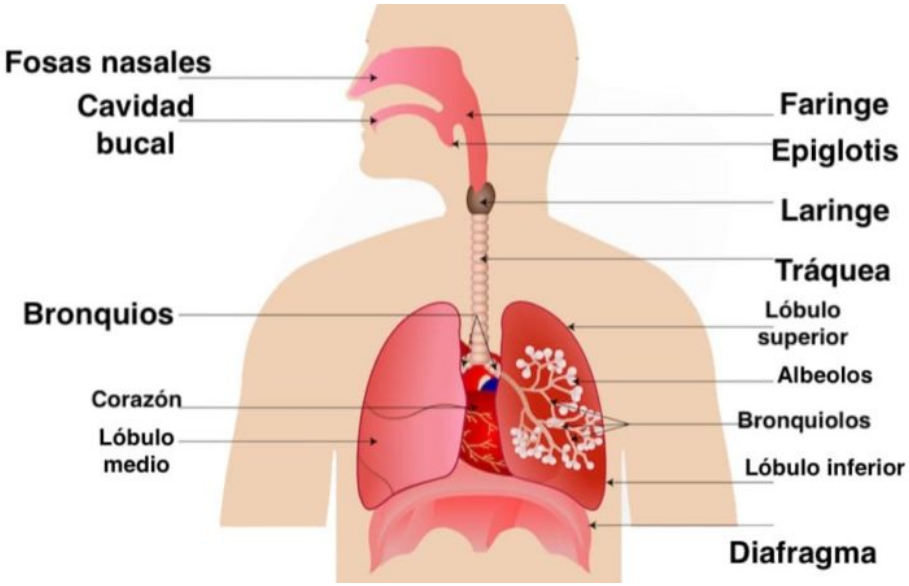


Figura 1: Aparato fonador
[23]

6.1.2. Ondas de sonido

Las ondas de sonido se transmiten a través del aire y la forma en la que se transportan genera cambios en la presión de este haciendo que varíe en relación a la presión atmosférica. Las ondas poseen diferentes características que las distinguen como lo pueden ser: la frecuencia, la amplitud o la velocidad. La velocidad del sonido depende del medio por el cual transcurre, en el aire está alrededor de los 343 m/s. El sonido al ser transmitido como ondas puede llegar a sufrir cancelación si se mezcla o suma con una onda igual que tiene un desfase

de 180° ya que al sumar ambas ondas de sonido estas se contrarrestan produciendo la cancelación lo cual dependiendo de la aplicación u objetivo puede ser beneficioso o perjudicial [24].

El volumen y el tono son atributos de gran relevancia no solo para el sonido en general sino también para la voz. El volumen es la presión que ejerce la onda sobre el medio, se mide en decibeles y se percibe como que tan fuerte es el sonido. La frecuencia determina el tono y esta se mide en *Hertz*. El oído humano no está hecho para poder escuchar todo el espectro de frecuencias que existen para el sonido. La voz humana se encuentra en el rango de quinientos a cuatro mil hertz por lo que aquellos sonidos dentro de este rango poseen una mejor receptividad al oído. El rango completo de audición es de veinte a veinte mil *Hertz* y con base en esta escala se categorizan los sonidos como infra, ultra o hiper sonidos [25].

Un sonido con un volumen alto se ve representado por una amplitud grande y un volumen bajo con una amplitud pequeña. La frecuencia por otro lado está relacionada al tiempo que transcurre entre cada ciclo de la onda. Una alta frecuencia produce sonidos agudos y significa que el tiempo entre valles o crestas de una onda es corto, una baja frecuencia produce sonidos graves y el tiempo es mayor. El principio de superposición permite que dos o más ondas sonoras se sumen algebraicamente para dar lugar a una onda donde todas estén representadas. El sonido que percibe el oído o se emite en un altavoz es una sumatoria de una gran cantidad de ondas con frecuencias y amplitudes distintas produciendo así una señal compleja. Tomando en cuenta los dos atributos y que cada onda sonora tiene una amplitud y frecuencia distinta a lo largo de la señal se aprecian zonas donde las ondas se superponen o se cancelan entre sí (Figura 2) [26].

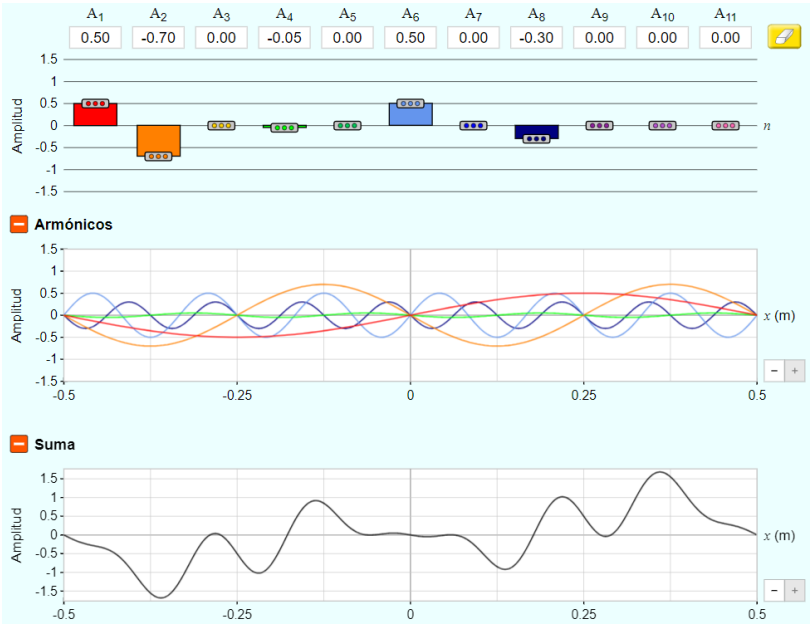


Figura 2: Principio de superposición de sumatoria de ondas [27]

6.1.3. Sonidos complejos

El lenguaje oral tiene como unidad mínima los fonemas, este tipo de sonidos son complejos si se analizan desde el punto de vista de las ondas sonoras pero elementales vistos desde la expresión oral, son la articulación mínima necesaria de un sonido vocálico o consonántico. Los fonemas forman parte de la fonología y para construir un lenguaje que permita la comunicación es necesario que cada palabra tenga asociada una forma fonológica con la cual diferenciarla de las demás al momento de la comprensión del habla. La gran cantidad de palabras que conforman un idioma se construyen con pocos fonemas ya que la forma de estos varía dependiendo de varios factores como los sonidos adyacentes, la posición en la sílaba, el tipo de sílaba, entre otros. Esto genera que cada fonema tenga una gran cantidad de variaciones donde suena diferente con base en todos estos factores [28].

El reconocimiento de las palabras no es posible sin los fonemas (Figura 3), más cuando existen tantas palabras similares que varían en pocas letras y sonidos, por ello los fonemas juegan un rol muy importante para poder diferenciar entre estas. Los fonemas son sonidos distintivos ya que permiten la identificación única, pero a su vez son abstractos ya que son aprendidos y es el sonido que el cerebro asocia con una letra o composición de letras para poder formar a partir de estas las palabras con las cuales nos comunicamos. En el idioma castellano existen 24 fonemas, cinco de ellos son vocálicos y los otros diecinueve consonánticos [29].

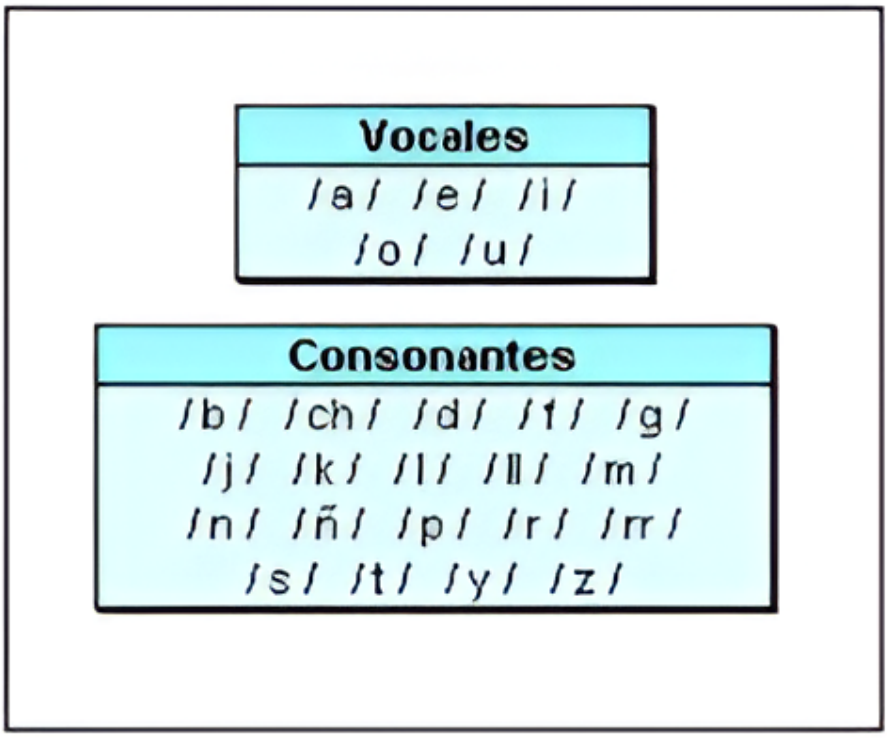


Figura 3: Fonemas vocálicos y consonánticos
[30]

6.1.4. Diferencias entre sujetos

La voz humana no es igual para todas las personas, cada uno tiene una forma particular dependiendo de su fisiología. La tesitura es la extensión de frecuencias de sonido que alcanza la voz de una persona, algunos tienen un mayor rango de frecuencias que otros. El color es el grosor de la voz y esto va más enfocado al tono promedio con el que habla la persona, una voz oscura es más gruesa y una clara es fina o aguda. Los hombres tienden a tener la voz más oscura que las mujeres y que los niños. Luego dependiendo del género y la edad existen distintas clasificaciones que son generales pero utilizadas mayoritariamente en la ópera. Las voces blancas corresponden a las voces de los niños antes de alcanzar la pubertad. Las femeninas se dividen de más aguda a más grave como soprano, mezzosoprano y contralto. Las masculinas como tenor, barítono y bajo (Figura 4) [31].

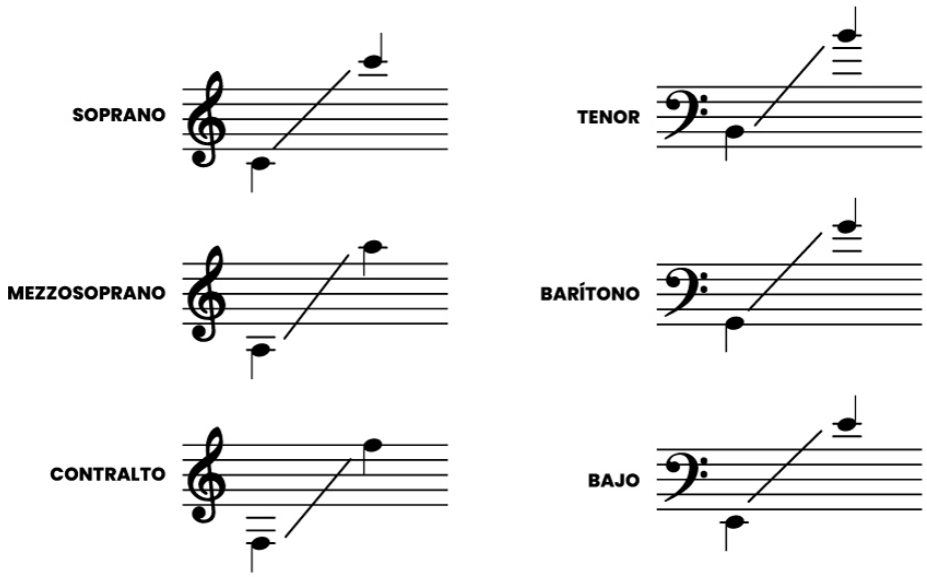


Figura 4: Rango de notas para cada tesitura de la voz correspondiente a su extensión de frecuencias [32]

6.2. Reconocimiento de voz

El reconocimiento de voz es la capacidad de mediante algoritmos y software transformar las ondas de sonido capturadas en el texto correspondiente. Estos sistemas son capaces de lograr esto al darle procesamiento a la señal que luego compara estos patrones del sonido con los fonemas para luego construir las palabras utilizando la similitud entre el habla reconocida y el amplio vocabulario de palabras que existen en el lenguaje analizado. La capacidad de computación actual hace posible que una gran cantidad de información analógica sea procesada en cuestión de segundos para obtener una transcripción de forma casi inmediata. El reconocimiento tiene aplicaciones dentro de muchos campos distintos. Uno de ellos es el de la medicina en el cual algunos ejemplos son conversaciones computadora-paciente para

agendar citas, solicitar medicina o chequeos básicos, interfaces de interacción con pacientes totalmente controladas por voz, en vez teclados y botones, transcripción de reportes dictados por doctores, etc [33].

6.2.1. Grabación del sonido

Para poder trabajar con la señal de sonido emitida por la voz es necesario transformarla en información que puede ser procesada por la computadora. Para esto se debe de discretizar la señal, esto se logra evaluando el estado de la señal en un intervalo de tiempo específico para poder transformar la señal continua en función del tiempo a una serie de datos de cómo varía la señal cada vez que es medida (1). A este proceso de transformación se le conoce como sampling, para no perder información de la señal análoga se utiliza el teorema de muestreo de Nyquist el cual indica que la frecuencia de muestreo debe de ser mayor al doble de la frecuencia más alta en la señal original (Figura 5). La frecuencia más grande en articulación del habla se encuentra alrededor de los 5.7 kHz por lo que para el muestreo se debe de utilizar una frecuencia mayor a 11.4 kHz. Una mayor frecuencia de muestreo también representa mayor espacio y recursos computacionales por lo que aumentar de sobremanera este valor puede llegar a ser contraproducente [34].

$$x[n] = a(nT) \tag{1}$$

Donde:

$$\begin{aligned} x[n] &: \textit{señal discreta} \\ a(nT) &: \textit{señal continua evaluada en los instantes de muestreo } nT \\ n &: \textit{índice discreto} \\ T &: \textit{período de muestreo} \end{aligned}$$

Luego de realizar el muestreo de la señal también se debe realizar un proceso de cuantización. Los valores extraídos de la señal real son transformados a una escala discreta asignándoles al valor más cercano dentro de los nuevos límites. La resolución es la cantidad de niveles a los cuales es posible asignar un valor mapeado dentro de los nuevos límites, esta cantidad está definida en bits y determina a su vez la forma que adquiere la gráfica en escalera donde se representa cada muestra. Cuando se realiza la cuantización, es importante tomar en cuenta que existe un error asociado el cual representa la diferencia que existe entre el valor muestreado y la cuantización de este valor mapeado. Este error suele ser cercano a $\frac{1}{2}$ del *step size* de cuantización. En señales de audio reducir este error es fundamental debido a que significa ruido añadido a la muestra y puede alterarla de manera significativa [36].

6.2.2. Procesamiento de las señales

Para poder reconocer las palabras que conforman la señal de voz ya digitalizada es necesario dividir el problema en pequeñas divisiones. Esto se logra por medio de *windowing* que es un método de procesamiento de señales en el cual la señal se analiza por segmentos

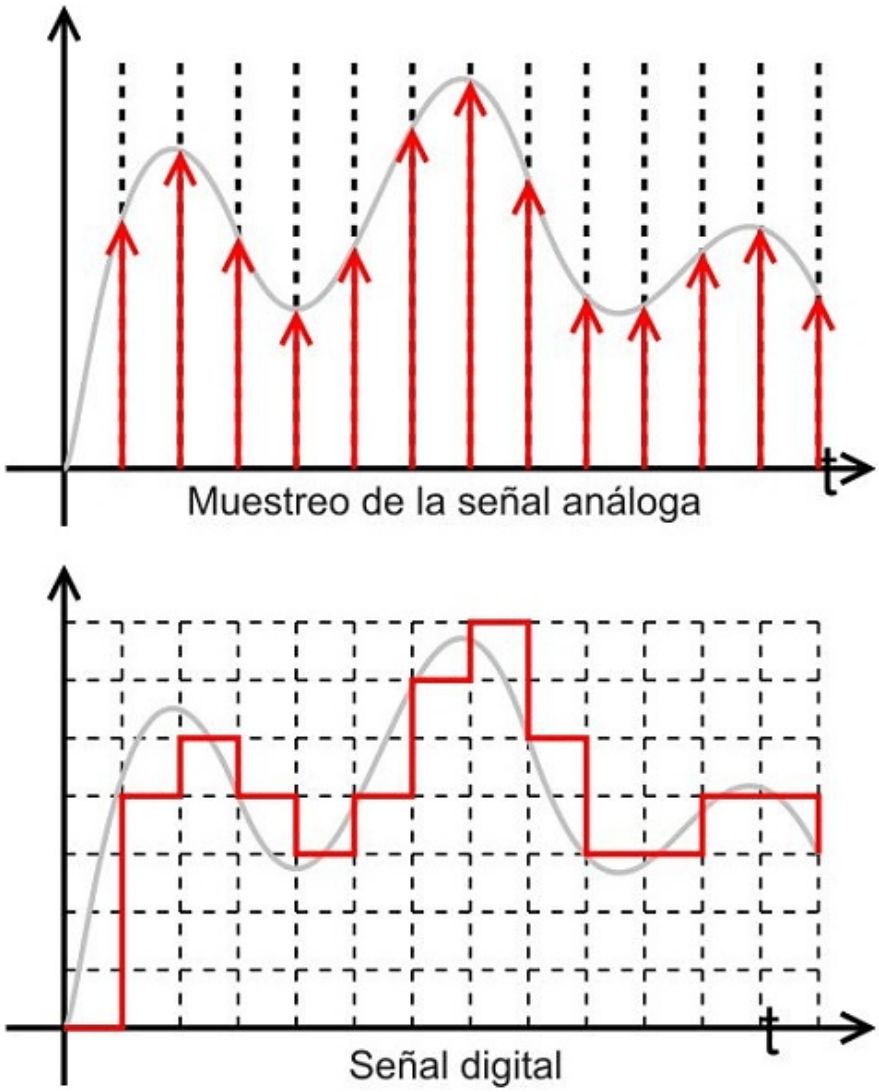


Figura 5: Señal digital producida del muestreo de la señal analógica [35]

utilizando una ventana temporal que es multiplicada por la señal convirtiendo los bordes en cero para enfocarse en la señal objetivo [37]. Esto es necesario ya que una señal de audio es continua por lo que para extraer sus partes elementales se debe aislar de lo demás. La elección de la ventana es importante ya que para el procesamiento posterior es importante evitar las discontinuidades. Por esta razón cuando se procesa audio una de las ventanas más utilizadas es la de Hamming [38].

El resultado de aplicar la ventana se conoce como marco o recuadro y se diseña para contener únicamente un fonema, esto se hace para analizar la señal fonema por fonema y así reconstruirla a partir del análisis de cada uno de estos recuadros. La ventana contiene un conjunto de diferentes espectros de frecuencia y para determinar los componentes de la señal se realiza análisis con la transformada discreta de Fourier. Se obtiene la magnitud y fase de cada componente frecuencial para ver cuáles son las principales frecuencias que componen la señal original que se analiza. Este proceso suele ser bastante extenso y complejo por lo que

se acostumbra a utilizar algoritmos de computación que realizan la transformada de manera interna como lo es la transformada rápida de Fourier (FFT) y luego se pueden extraer los resultados para su posterior análisis [38].

Antes de formar las letras, los morfemas y las palabras es necesario recopilar toda la información que se procesó con la transformada de Fourier. Esta información se ingresa al conjunto de encoder-decoder como ‘’n” cantidad de vectores acústicos de la misma duración de tiempo donde cada vector corresponde al marco analizado previamente. El resultado o salida de este conjunto viene a ser las letras o palabras que corresponden a cada uno de los vectores ingresados [38].

6.2.3. Algoritmos de reconocimiento

Los algoritmos que utilizan *machine learning* siguen todos un patrón bastante similar de procesos previo al entrenamiento para la predicción. Todo comienza con una selección de bases de datos, la cual debe contener un alto número de ejemplares ya que en esto se basa el *machine learning*. Para reconocimiento de voz es recomendable utilizar archivos de tipo .wav porque abarcan todo el espectro de frecuencias audibles. Luego de esto se limpia la señal, usualmente utilizando ventanas Hamming (Figura 6). Continúa el preprocesamiento con el filtrado de la señal, un filtro utilizado con regularidad es el filtro de preénfasis (2).

$$y(n) = x(n) - a \cdot x(n-1) \tag{2}$$

Donde:

$y(n)$: *señal discreta de salida del filtro de preénfasis*
$x(n)$: *señal discreta de entrada original*
$x(n-1)$: *señal de entrada en el instante discreto anterior*
a : *coeficiente de preénfasis, normalmente utilizando valores de* $0.9 < a < 1.0$

Luego un análisis de Fourier utilizando la FFT para pasar del dominio del tiempo al de frecuencia. Se transforman a vectores manejables de la señal subdividida y están listos para utilizarse en clasificadores de *machine learning* [40].

A partir de este punto cada método es diferente, algunos de ellos se basan en listas de decisión donde un elemento es asignado a una categoría dependiendo a qué regla se asemeje más. Otros como el Naive Bayes utilizan teoremas de probabilidad como el teorema de Bayes. Los modelos de *decision trees* o más complejos como *random forest* construyen los árboles y asignan elementos con base en estas cadenas de decisiones. Sin importar el modelo que se utilice es indispensable que estos cuenten con tres fases: entrenamiento, prueba y evaluación del algoritmo. El entrenamiento es la parte más importante debido a que es cuando el algoritmo aprende a generar patrones y hacer predicciones con base en información real que se asemeja a los elementos con los que trabajará. Luego la fase de prueba es para descubrir cómo se comporta el algoritmo posterior a su entrenamiento, al trabajar con elementos de

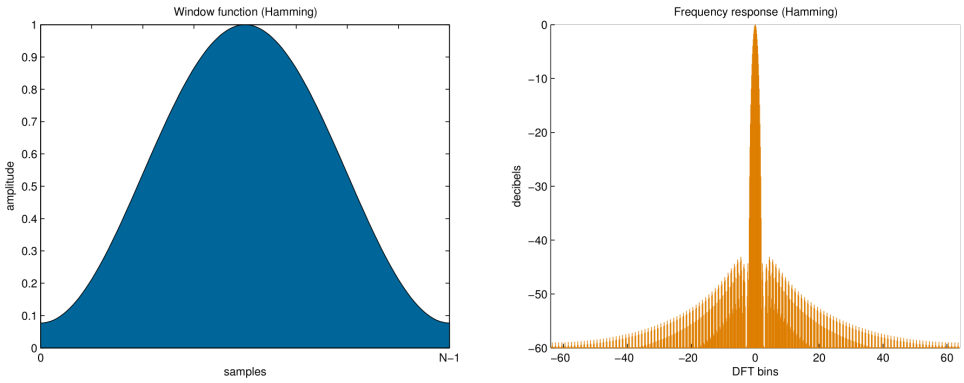


Figura 6: Ventana de Hamming
[39]

audio es de suma importancia que los elementos hayan sido tratados con el mismo preprocesamiento que los de entrenamiento para evitar discordancias en los resultados. La última parte es verificar con métricas de exactitud, precisión, especificidad, sensibilidad, etc., qué tan bien predice el algoritmo [40].

Parte importante del reconocimiento de voz es filtrar todo el ruido de la señal objetivo. Esto se logra por medio de algoritmos de realce de la voz. Tres de los más comunes son el algoritmo iterativo de mejora de señal, algoritmo de subespacio basado en la mejora del habla y la sustracción espectral no lineal. El primero realiza varias iteraciones del mismo proceso de filtrado para eliminar en su mayoría el ruido. La cantidad de iteraciones no es fija sino que varía dependiendo de la cantidad de ruido, después de cada iteración analiza el ruido residual y repite el proceso hasta alcanzar un nivel aceptable. El método de subespacios se utiliza en otras aplicaciones, no solo en reconocimiento de voz. Este método utiliza eigendescomposiciones de matrices para determinar los componentes frecuenciales, aísla el subespacio que corresponde al ruido y lo elimina de la señal. El último es de los más básicos y famosos, está basado en factores de sustracción por cada banda de frecuencia y FFT para diferenciar las señales de ruido del habla [41].

6.2.4. Ejemplos de usos cotidianos

Este tipo de tecnología y algoritmos no se utiliza únicamente en proyectos relacionados al área de salud o a nivel militar, se encuentra cada vez más inmersa en la vida cotidiana. Ejemplo de esto son los asistentes inteligentes por voz (Figura 7). Es común encontrar en un hogar uno de estos dispositivos. Dentro de sus principales funciones se encuentran reproducir música, configurar alarmas, entablar conversaciones, buscar información o configurar recordatorios y tareas [42]. Un ejemplo de funcionamiento de estos asistentes es el Amazon Echo Dot, que utiliza aplicaciones de Android para su funcionamiento lógico. Divide su reconocimiento de voz en dos ya que utiliza modelos básicos de comparación para las detecciones iniciales de llamada al convertir a texto y comparar con expresiones definidas. Mientras que el reconocimiento de las interacciones más específicas emplea algoritmos más complejos donde transcribe la petición y analiza la información para realizar una tarea relacionada [43].

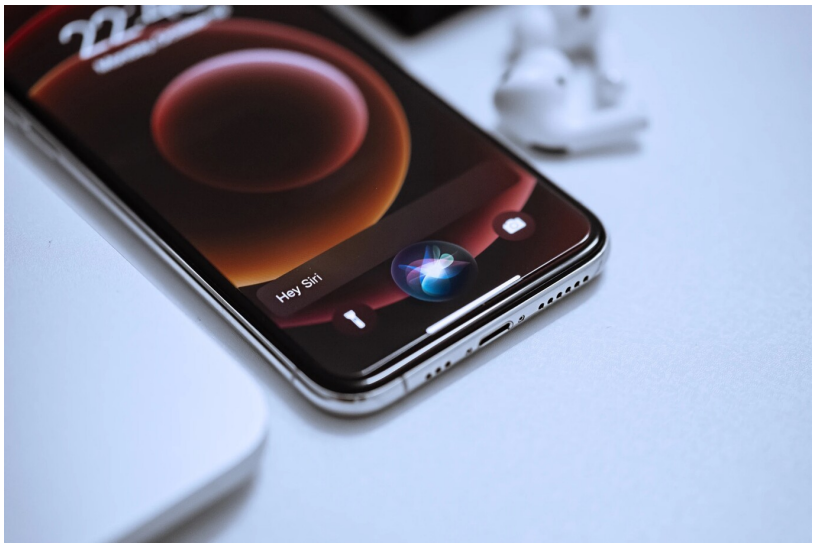


Figura 7: Asistente de voz de sistema operativo iOS (Siri)
[44]

Otro ejemplo que irrumpe con fuerza en la vida de las personas son las aplicaciones de casas inteligentes que suelen venir acompañadas de controladores por voz. Muchas de estas aplicaciones utilizan interfaces de programación de aplicaciones (API) para implementar reconocimiento de voz para la selección de comandos. Estas interfaces no solo incluyen los algoritmos de síntesis de la voz sino también el procesamiento del lenguaje natural para generar acciones sobre el sistema o plataforma. Suelen identificar y utilizar los diferentes sistemas de reconocimiento del sistema operativo presente como lo puede ser Cortana en Windows o Siri en iPhone por dar algunos ejemplos. Los sets de entrenamiento de estas aplicaciones están enfocados en los posibles escenarios a los que se puede ver sujeto, son específicos de cuestiones como el clima, la temperatura, preguntas generales, comandos o humedad. Para cada algoritmo es necesario fijar cuales son los escenarios objetivo para mejorar sus capacidades de comprensión en esa área [45].

6.3. Transcripción de voz

La transcripción de la voz es la transformación de toda la información recopilada al momento de una conversación o dictado a texto que puede utilizarse posteriormente. Esto sirve para mantener el registro o para analizar de manera más detallada la información. La transcripción en muchos casos se realiza de manera manual por un profesional capacitado, pero requiere de tiempo, por esta razón existen transcriptores de voz automáticos. Los algoritmos de transcripción automática más complejos utilizan inteligencia artificial y procesamiento del lenguaje natural para transformar el habla a texto [46].

La transcripción humana suele tener tiempos de duración en el rango de días mientras que la automatizada se completa en minutos, un archivo de audio que dura treinta minutos se puede transcribir en menos de cinco minutos. Otra de las ventajas que presenta este tipo de tecnología con relación a los métodos convencionales es la privacidad y seguridad

que ofrecen. Cuando la transcripción se hace de forma automática con un dispositivo no se necesita a un tercero que intervenga y tenga acceso a la información en cuestión. Todo el proceso se lleva a cabo de manera computarizada. Este tipo de tecnología no es perfecta pero cuando se trabaja con archivos de audio donde la articulación de las palabras, la claridad y calidad del audio son las adecuadas se alcanzan niveles de precisión superiores al 95 % [47].

6.3.1. Procesamiento del lenguaje natural

Los algoritmos de PLN sirven para analizar datos sin un orden o estructura definida y es capaz de producir modelos de predicción para dotarlos de estas características a partir de ciertas reglas adoptadas del lenguaje. Une los algoritmos de aprendizaje profundo con la lingüística computacional. La lingüística computacional es la que utiliza la ciencia de datos para analizar todos los conceptos y fundamentos del lenguaje humano y el habla. En esta ciencia se estudia tanto el significado de las palabras como el lugar que ocupa cada palabra dentro de la estructura de las oraciones [48].

Cuando se realiza el análisis de las palabras se puede llevar a cabo de dos maneras distintas. Para el reconocimiento y transcripción de voz el más relevante es el análisis de circunscripción, este se realiza construyendo un árbol sintáctico que utiliza la raíz sintáctica como guía para determinar las cadenas de palabras, este análisis es el más complejo para alcanzar un resultado que sea entendible y tenga sentido. El otro tipo de análisis se conoce como análisis de dependencia se fija en la relación que existe entre las palabras, entender la relación que tienen permite determinar la manera en la que deben de ser ordenadas la identificación de los diferentes componentes de una oración [48].

Como en otros tipos de inteligencia artificial, los modelos auto supervisados producen los resultados más confiables y precisos pero este tipo de modelos es demasiado costoso en cuestiones de recursos y tiempo. Por esta razón el PLN se suele implementar con otras perspectivas, ejemplos de estas son: PNL basado en reglas, de aprendizaje profundo y estadístico [48].

El primero está basado únicamente en árboles de decisiones donde se especifican las reglas condicionales y no hace uso ni de aprendizaje profundo ni tiene capacidad de inteligencia artificial por lo que es bastante básico y su capacidad de respuesta es bastante limitada. En el otro extremo están los de aprendizaje profundo que son los más recientes, utilizados y completos en la última época. Estos emplean una gran cantidad de datos no estructurados que entrenan al modelo de red neuronal. El estadístico se considera un término intermedio entre estos dos, este asigna una probabilidad de significado a cada uno de los elementos evaluados. Este modelo mapea elementos lingüísticos a vectores que se pueden manejar a nivel matemático y estadístico [48].

6.3.2. Limitaciones

Las diferencias en la pronunciación, acentos y demás factores que varían entre cada hablante representa una de las principales limitaciones a la hora de trabajar con reconocimiento y transcripción de la voz, sobre todo a la hora de entrenar un algoritmo. En el área médica el

manejo de la información proporcionada durante el reconocimiento sobre todo si es personal representa una serie de datos que requieren un alto nivel de protección de la privacidad por lo tanto es necesario tomar esto en cuenta siempre que se trabaja con este tipo de tecnologías en un escenario así. Muchos de estos algoritmos sólo se analizan desde el punto de vista de la efectividad de transcripción con respecto a la grabación, pero en este sector es de vital importancia desarrollar métodos que permitan analizar su efectividad en escenarios médicos reales donde la predicción con base en el resultado de transcripción sea evaluada [4].

6.3.3. Modelos de clasificación de palabras

En este campo existen diferentes modelos de clasificación que permiten transformar este tipo de información a palabras. Dos de los más comunes son el Connectionist Temporal Classification (CTC) Algorithm y los modelos ocultos de Markov. No todos los modelos de clasificación pueden ser utilizados para este tipo de aplicación porque la mayoría están acostumbrados a manejar elementos donde no existe la variación temporal que sí existe cuando se trabaja con fonemas o palabras donde existen personas que acentúan en diferentes partes.

El algoritmo de CTC trabaja específicamente con problemas donde no es posible mapear de manera exacta desde la entrada hasta la salida, con problemas para transformar un tipo de secuencias a otro. Este es un método de clasificación y ordenamiento que utiliza redes neuronales recurrentes sin necesidad de utilizar elementos segmentados anteriormente para el entrenamiento. Esto lo logra determinando una probabilidad de ordenamiento sobre todas las posibilidades de clasificación para una sola entrada. Utiliza un set de datos de entrenamiento para realizar un clasificador temporal de información que no ha visto antes, los resultados obtenidos de la red se convierten en una distribución condicional de probabilidad y sobre estas probabilidades se construye el clasificador que determina la secuencia más probable con base en la entrada [49].

Los modelos de Markov ocultos se derivan de las cadenas de Markov y es uno de los modelos probabilísticos más utilizados. Está definido por un conjunto finito de estados donde cada uno cuenta con una probabilidad asociada y se maneja una matriz de transición donde quedan especificadas los cambios entre estado y sus probabilidades de transición. Es posible ver lo que está sucediendo en el exterior del modelo, pero los estados en todo momento permanecen ocultos. Los cinco elementos principales para construir un modelo de Markov oculto son: los estados, la matriz de transición, los símbolos o elementos que se observan, la probabilidad de cada uno de los símbolos para cada estado y la probabilidad inicial de cada estado. Este modelo se ayuda de otros algoritmos como Viterbi para encontrar dentro de las secuencias con la mayor probabilidad aquella que más se acerque a un resultado real o el algoritmo de Baum-Welch que se utiliza para aproximar los parámetros de probabilidad del modelo a la realidad de las observaciones, se utiliza para mejorar el aprendizaje de este (Figura 8) [50].

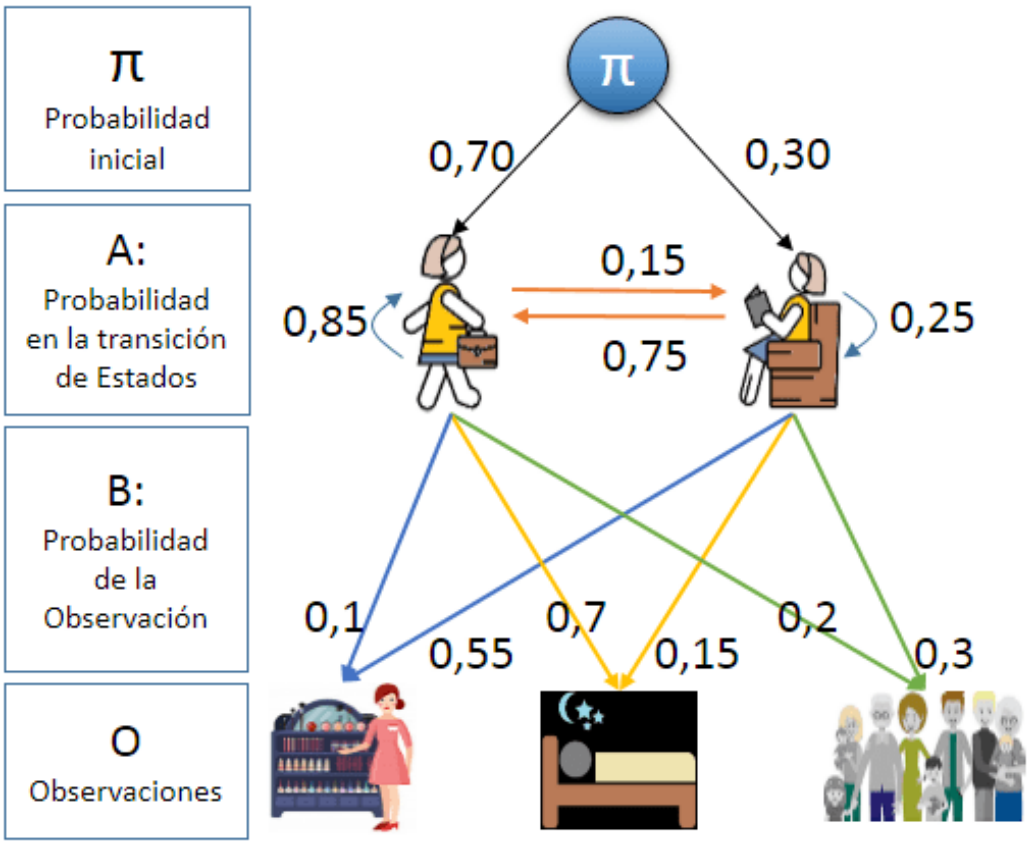


Figura 8: Ejemplo de modelo de Markov
[51]

6.4. Entrevistas médicas

La entrevista inicial permite al médico identificar todos los asuntos que pueden significar algún inconveniente para el paciente, tanto los evidentes como los sutiles. Los datos que se incluyen dentro de la entrevista deben ir enfocados en la perspectiva del paciente, no del médico o entrevistador. Es importante que en todo momento el paciente se sienta cómodo tanto a nivel físico como con su entorno porque en muchos casos la apertura a expresar todos los factores involucrados en su malestar se ve afectada por esto [52].

Al momento de abordar cuestiones que pueden llegar a ser sensibles para el paciente es importante tomar las debidas precauciones y consideraciones. Se debe garantizar la confidencialidad de la información que se está manejando, ser directo sobre la pregunta y evitar pedir disculpas por la pregunta, no presionarlo a emitir una respuesta y evitar el enfrentamiento. Es importante conocer los momentos y el tipo de preguntas donde se busca una respuesta corta o elaborada, si es necesario guiar al paciente sobre lo que se espera de su respuesta o permitirle que se exprese libremente para obtener la mayor cantidad de detalles posibles [52].

6.4.1. La metodología de la entrevista médica

Existe una estructura definida de qué partes constituyen una entrevista, pero siempre se deben de adaptar dependiendo de las circunstancias de cada paciente. Lo primero es completar el motivo de consulta, se debe de responder la pregunta *¿Por qué está aquí?*, se realizan preguntas sobre la duración de la complicación y si existen otro tipo de problemas que lo aquejan ya que el motivo principal puede derivarse de otros problemas que terminan siendo más alarmantes. Se continua con los antecedentes de la enfermedad, estos deben de encontrarse ordenados cronológicamente, especificar el estado de salud previo al problema, los síntomas con una descripción de su evolución, el impacto que ha tenido sobre su vida, el uso de terapias o medicamentos, etc [52].

Los antecedentes médicos sirven como una evaluación y valoración de referencia para la situación médica actual. Se consulta si han padecido enfermedades en su infancia, si tienen alguna enfermedad crónica, las vacunas con las que cuenta, si se ha realizado algún procedimiento u operación, alergias o medicinas que consume de manera regular. Los siguientes antecedentes por anotar son los familiares. Aquellos parientes consanguíneos de la familia inmediata o extendida que sufran de alguna enfermedad que pueda llegar a estar asociada al cuadro que presenta el paciente deben ser tomados muy en cuenta [52].

Es importante determinar el estado de salud de estos familiares y, si aplica, la causa de muerte de los familiares de primer grado, luego se realiza el mismo procedimiento con los de segundo y tercer grado. No todas las enfermedades son de relevancia, las de mayor interés son cáncer, accidentes cardiovasculares, diabetes, epilepsia, enfermedades sanguíneas, etc. Por último, se tratan los antecedentes familiares y sociales, estos incluyen diferentes tipos de información relacionada a su ambiente y conducta. Inicia con los datos personales donde se indaga acerca del entorno actual y pasado de la persona. Luego los hábitos, estos incluyen la frecuencia del ejercicio, consumo de drogas. Las condiciones domésticas y entorno dan una idea sobre la forma en la que vive la persona, su estado económico y los factores que rodean su vida diaria que pueden tener un impacto sobre su salud. También se pregunta sobre la ocupación debido al efecto que pueden tener las condiciones y tiempo que pasa en el trabajo. Otro tema que se aborda en esta sección es la preferencia religiosa ya que esta puede estar relacionada con la alimentación o la prohibición de acceso a atenciones médicas como lo puede ser la recepción de sangre [52].

6.4.2. Revisión preliminar de sistemas

Además de la entrevista sobre el motivo y los antecedentes el médico debe de llevar a cabo una revisión de sistemas (Figura 9). Empieza por los síntomas generales que incluyen fiebre, malestar, fatiga, dolor y demás. Cambios en la apariencia, textura o coloración del pelo, uñas o piel. Luego se realiza un análisis de la cabeza y el cuello, esta se divide en 5 secciones: una general, los ojos, oídos, nariz y garganta. De forma general se pregunta si se han sufrido conmociones, lesiones o mareos, en los ojos es necesario analizar la agudez visual, glaucoma, borrosidad o cambios en la capacidad de la vista, en los oídos se busca la presencia de dolor, infecciones o vértigo, en la nariz se analiza si existen obstrucciones, secreciones o dolor y en la garganta inflamaciones, abscesos o llagas [52].

Así también, la evaluación de los ganglios linfáticos, el tórax y los pulmones donde puede existir dolor al momento de respirar, sibilancias o tos. En el corazón hay que desestimar dolores en la zona, las causas y características del dolor, presencia de hipertensión, etc. En la sangre se buscan hematomas, anemias u otras anomalías; en el sistema gastrointestinal la condición de su digestión, si presenta cambios en el apetito, náuseas, vómitos, coloración distinta en las heces, diarrea u otras condiciones que puedan estar relacionadas. Para el músculo esquelético se estudia si hay restricciones del movimiento o rigidez en alguna de las articulaciones, deformaciones a nivel óseo o enrojecimiento en alguna zona específica. A nivel neurológico y psiquiátrico se evalúa si ha sufrido convulsiones, temblores, pérdida de la memoria, depresión, ansiedad, trastornos del sueño, fallas de coordinación o pensamientos suicidas [52].

Es importante que en la revisión de sistemas se dé un enfoque especial a aquellos que están conectados con el motivo de la consulta ya que no siempre todos tienen la misma relevancia. De igual forma hay factores del paciente que afectan tanto la forma en la que se lleva a cabo la entrevista como la revisión de sistemas como lo son el género y la edad. No es lo mismo examinar a un niño que a alguien de la tercera edad y las evaluaciones en mujeres no son las mismas que en los hombres. Para la revisión de sistemas en mujeres se debe de estudiar la edad de la menarquia, regularidad de las menstruaciones, autoexamen de mamas, frecuencia de las relaciones sexuales, infertilidad, cantidad de gestaciones, si ha tenido algún aborto, la duración de los embarazos y si existió algún tipo de complicación en estos. Mientras que, para los hombres se consulta sobre el inicio de la pubertad, dolores en los testículos, infertilidad, líbido, impotencia o problemas a la hora de tener relaciones sexuales [52].

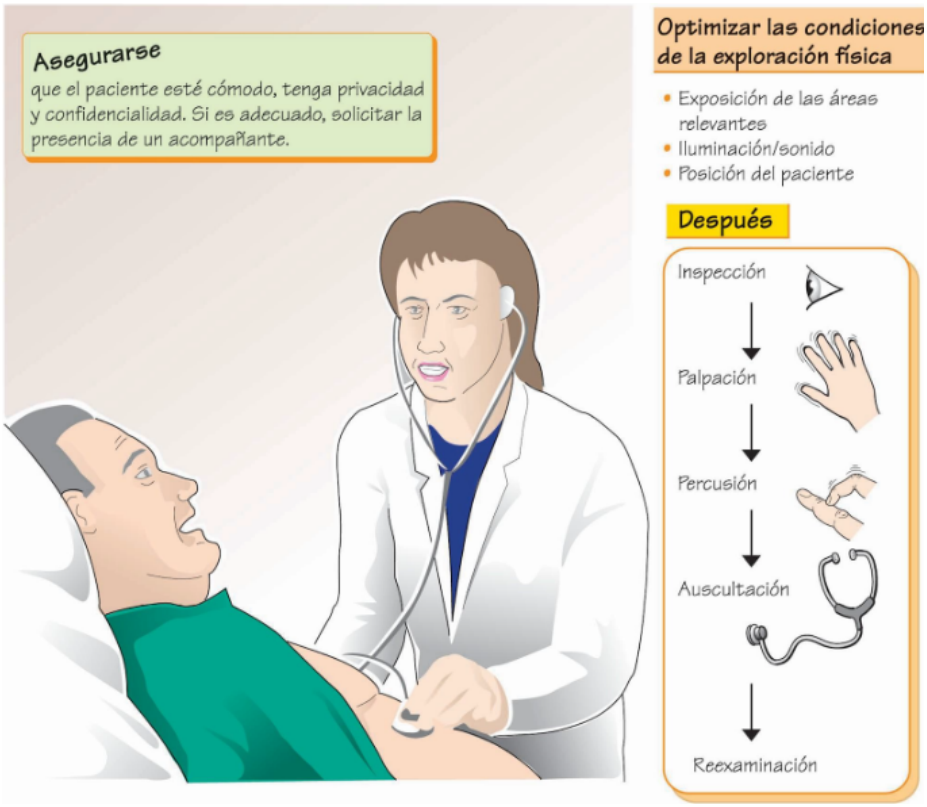


Figura 9: Protocolo de revisión preliminar de sistemas
[53]

CAPÍTULO 7

Metodología

7.1. Esquema de entrevista

7.1.1. Reuniones con médicos

Se realizan dos reuniones distintas con personal del área de salud para comprender de mejor manera como se llevan a cabo las entrevistas médicas, el protocolo que se sigue y los distintos apartados que la conforman. En estas reuniones se aborda la importancia de la anamnesis, los factores principales que determinan la especialidad a la que debe de ser asignado un paciente, las secciones de la historia clínica que tienen un mayor peso a la hora de poder realizar esta clasificación y el tipo de preguntas útiles para referir a una especialidad. Además de esto también se consulta a los médicos sobre recomendaciones que se deben tener en cuenta al momento de llevar a cabo la entrevista preliminar como los listados de patologías, las zonas del cuerpo relacionadas a la complicación que experimenta el paciente, los antecedentes que pueden tener un rol importante. Todo esto sirve de base para generar las encuestas que determinan las preguntas a utilizar para la entrevista.

7.1.2. Encuestas a personal médico

Para determinar las preguntas que se incluyen en el protocolo de la entrevista dentro del algoritmo se realizan encuestas a diferentes miembros que forman parte del personal médico en diferentes hospitales dentro del país. Esta información se recopila utilizando un formulario de Google Forms (Google Inc., CA, USA), las entradas son de manera anónima ya que los datos personales no son necesarios en este caso. Dentro del formulario se informa sobre el propósito de la encuesta, la anonimización de sus respuestas y el canal de comunicación en caso de presentar dudas con respecto al manejo de la información.

Dentro de la encuesta se solicita que indique el puesto o rol que funge dentro del hospital y luego se procede a explicar detalladamente las instrucciones sobre las respuestas que se esperan en la siguiente sección. Se indica al encuestado que considere el caso en el que solo tuviese la posibilidad de realizar diez preguntas a un paciente para determinar la especialidad a la que debería de ser asignado. Luego encuentra un espacio para cada una de sus preguntas las cuales se pide que no sean demasiado generales o abiertas produciendo que toda la afección sea descrita en una sola respuesta, sino más bien complementarias entre sí (Figura 10). Al realizar el análisis numérico de la cantidad de repeticiones por tipo de pregunta se determinaron aquellas con mayor presencia dentro de la muestra. Basado en esta estadística y la opinión experta de un médico, se determinó el esquema final de preguntas para la entrevista.

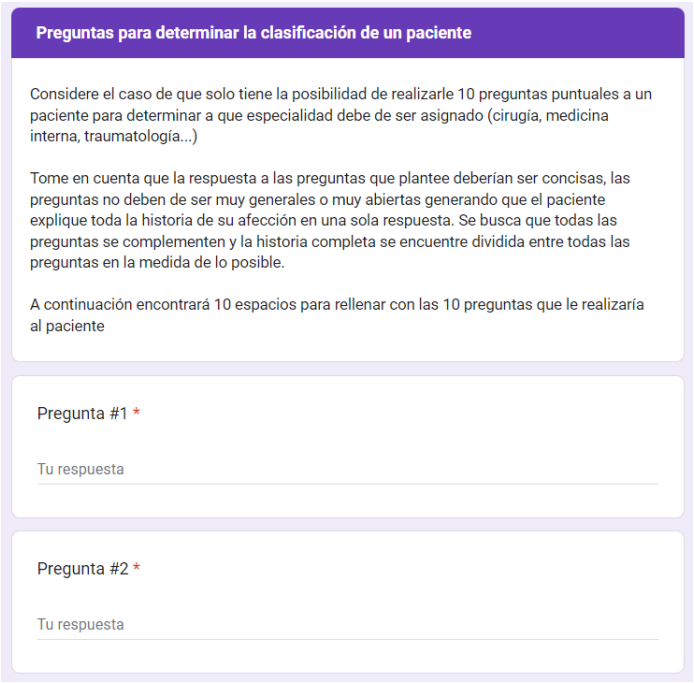


Figura 10: Encuesta realizada por personal médico para determinar esquema de entrevista

7.2. Algoritmos de reconocimiento

7.2.1. Versiones del algoritmo

Se programan diferentes versiones del algoritmo de reconocimiento (Figura 11) incluyendo en cada nueva versión una característica que busca mejorar el algoritmo consiguiendo aumentar la efectividad de transcripción del habla. El algoritmo se crea utilizando la plataforma de desarrollo Visual Studio (Microsoft, WA, USA) con la interfaz visual de Windows Forms Application y programado utilizando lenguaje C (Microsoft, WA, USA) orientado a objetos, esto con el fin de crear una interfaz visual con la cual llevar cabo la entrevista. La primera versión es la más básica donde solo se programa para que identifique cuando el micrófono está recibiendo una señal de audio y transcribe toda la señal de audio mientras el

micrófono percibe una entrada. Al motor de reconocimiento se le especifica la información de cultura que es el parámetro que le permite saber a qué idioma pertenecen las palabras que estará reconociendo cuando se carga la gramática de dictado que en este caso es el español.

Cada una de las versiones se construye sobre la anterior o añade elementos de grabación a las ya existentes para evaluar qué cambio generan sobre su efectividad de transcripción. El algoritmo que le sigue incluye un diccionario ampliado que se implementa sobre el constructor de la gramática ya existente del idioma español, esto se hace para dar énfasis en el tipo de palabras que se esperan en una entrevista médica y reducir las posibilidades de que esas mismas palabras sean confundidas por otras similares al momento de la transcripción. Estas palabras fueron extraídas de ejemplos de fichas clínicas y enterados de casos clínicos de estudiantes de medicina de la Universidad Rafael Landívar. El diccionario ampliado se divide en:

37 partes externas del cuerpo (codo, boca, ojos, rodilla, etc.)

41 órganos (útero, apéndice, ligamento, intestino, etc.)

54 condiciones, signos o síntomas (cansancio, sarpullido, sordera, estrés, etc.)

40 términos complementarios (tumor, diálisis, infección, punzante, etc.)

La penúltima versión además de contar con el diccionario ampliado cambia la entrada del motor de reconocimiento para poder realizar procesamiento con la señal. Este algoritmo utiliza dos botones de inicio y parada para grabar la señal de audio (Figura 12), en este caso la voz. Luego esta señal se almacena en un archivo .wav (waveform audio file format) con un formato de 44.1 kHz de frecuencia de muestreo, 16 *bits* por muestra y un único canal. Al concluir con la grabación y almacenamiento del archivo inmediatamente se da el procesamiento de la señal. Para el procesamiento se discretiza la señal y se construye un filtro pasabandas Butterworth virtual de grado 5 con frecuencias de corte de 70 y 8000 Hz que se aplica a la señal cruda 4 veces consecutivas para reducir el espectro de frecuencias de la señal filtrada al espectro de la voz y sus fonemas. Esta señal se almacena en un archivo .wav nuevo con el mismo formato y se utiliza como entrada del motor de reconocimiento que transcribe la señal entera para desplegar el texto reconocido.

Al evaluar las versiones ya mencionadas se toma la decisión de probar una versión que utiliza un motor de reconocimiento diferente a las anteriores. Este motor no es local, no realiza el reconocimiento en el mismo dispositivo donde se hace el procesamiento de la señal ya que utiliza un servidor en la nube. Se crea la última versión que sigue captando la señal, almacenándola en un archivo .wav, procesando la señal por medio del filtro pasabandas y re almacenándolo en otro archivo ya procesado. La diferencia es que luego se traslada a los servidores de transcripción en tiempo real basados en inteligencia artificial utilizando una Interfaz de Programación de Aplicaciones de la empresa Assembly AI (Assembly AI Inc., CA, USA) que se integra al código y cuando termina de procesarlo regresa el texto transcrito para ser desplegado en la interfaz local.

Figura 11: Evolución de las versiones de algoritmo de reconocimiento de voz evaluadas

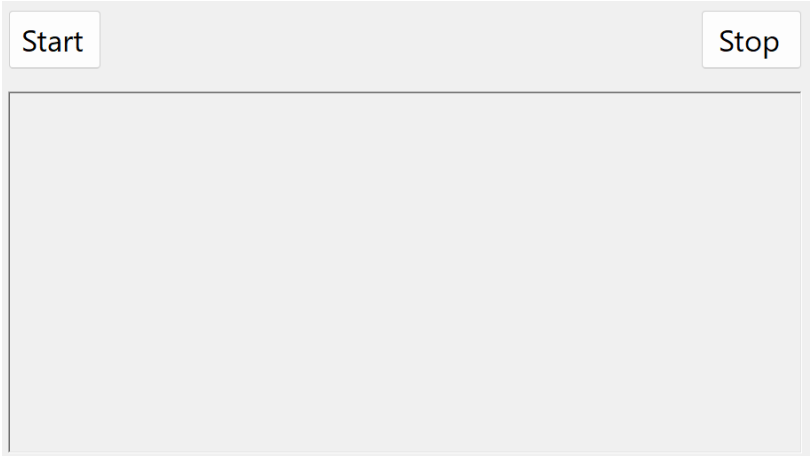


Figura 12: Interfaz gráfica para la versión con grabación y almacenamiento de archivo de audio .wav

7.2.2. Pruebas de reconocimiento

Para verificar la efectividad de reconocimiento de cada uno de los algoritmos, se realizan pruebas (Figura 13) con diferentes individuos de ambos géneros y se analizan las métricas obtenidas de cada uno. Estas son llevadas a cabo con ocho individuos, cuatro hombres y cuatro mujeres, sin restricción de edad. Previo a la realización de la prueba todos los participantes firman un consentimiento informado y se les explica de qué va a constar el procedimiento y cuáles serán los datos utilizados para la investigación. Se realizan en un espacio de grabación para evitar la interferencia de ruidos externos que afecten los resultados.

Figura 13: Resumen de las pruebas realizadas para verificar la efectividad de reconocimiento de voz de los algoritmos

Se les solicita a los participantes que lean detenidamente las frases que estarán recitando en voz alta durante las pruebas, esto les permite familiarizarse con ellas y reducir las confusiones o equivocaciones al momento del reconocimiento. El total del texto a leer por los individuos se constituye por tres secciones de cinco frases cada una, para un total de quince. Las tres secciones que conforman el texto son: frases relacionadas al sujeto como decir cuál es el nombre, donde estudia, la edad que tiene; frases relacionadas a una entrevista médica como el medicamento que consume o síntomas y por último frases de un contexto diferente como literatura, poesía o películas. Para todos los participantes las frases son las mismas, no varían dependiendo de su información personal. Las quince frases a utilizar son:

1. *Mi nombre es Julián Monterroso.*
2. *Estudio Matemática en la Universidad del Valle de Guatemala y actualmente estoy cursando el cuarto año.*
3. *Tengo 23 años y, por las tardes, suelo salir a caminar con mis perros.*
4. *Mi hermano tiene alergia a las nueces y a las fresas.*
5. *En mis tiempos libres, suelo leer algún libro y hacer deporte cerca de mi casa.*
6. *Ayer tuve un accidente de moto, tengo bastantes moretones, me rompí el brazo y el fémur.*

7. *Vengo porque llevo ya un mes con los ojos rojos como si se me reventaran los vasos, además de que he tenido muchas náuseas repentinas.*

8. *En la noche, me dan taquicardias y hace una hora sentí un fuerte dolor en el pecho, cerca del corazón, como si me estuvieran apuñalando.*

9. *El doctor me recetó una medicina que era para eliminar bacterias dentro de mi intestino delgado.*

10. *Tuve unos exámenes de sangre ayer porque necesitaba donar sangre para un amigo.*

11. *La obstrucción de las vías respiratorias por un cuerpo extraño provoca una asfixia repentina, si no se resuelve provoca una hipoxia grave.*

12. *Más vale pájaro en mano que 100 volando y camarón que se duerme se lo lleva la corriente.*

13. *La vida no es esperar a que pase la tormenta; es aprender a bailar bajo la lluvia*

14. *Yo no hablo de venganzas ni de perdones; el olvido es la única venganza y el único perdón. Jorge Luis Borges*

15. *Es en las noches de diciembre, cuando el termómetro está a cero, cuando más pensamos en el sol. Los Miserables*

A cada uno de los algoritmos se les agrega una función que extrae todo el texto que se encuentra en el recuadro de texto enriquecido y lo almacena en un archivo de texto .txt. La prueba es la misma para las tres versiones del algoritmo. El individuo lee cada frase y realiza una pausa, se espera a que el programa realice la transcripción y luego se le indica que prosiga con la siguiente frase. Las primeras tres versiones se analizan utilizando el micrófono que trae integrado la computadora Lenovo modelo Ideapad 5-15IIL05 Type 81YK (Lenovo Group, Pekín, China) y las últimas dos versiones se guardan el audio en un archivo y realizan el filtrado de la señal antes de la transcripción también se analiza utilizando un micrófono profesional M-Audio's Uber Mic (M-Audio, RI, USA). Todo este experimento tiene una duración aproximada de quince a veinte minutos para completarse con cada persona.

7.2.3. Pruebas al algoritmo elegido

Al analizar la efectividad de transcripción de todas las versiones creadas se elige la que presente la media de efectividad mayor para poder asegurar que la versión a utilizar en el programa final tendrá la menor cantidad de errores a la hora de la transcripción. En este caso esta versión es la última creada, la que utiliza inteligencia artificial para transcribir el audio enviado al servidor en la nube. Al contar con el algoritmo elegido se realizan dos pruebas fundamentales para determinar que tan viable es utilizarlo dentro de la versión final y analizar la forma en la que funciona entorno a diferentes situaciones o escenarios.

La primer prueba es la de velocidad de transcripción o velocidad de respuesta. Para esta prueba se realiza la transcripción de las mismas respuestas a una entrevista repitiendo esto 10 veces. Dentro de estas respuestas se encuentran textos de una sola palabra como lo puede

ser Sí o No y también respuestas mucho más complejas de hasta 18 palabras. Con esto se obtiene una muestra que refleja el tiempo que tarda en transcribirse un texto por cantidad de palabras. Para medir el tiempo de respuesta se cronometra el tiempo que transcurre desde el momento en el que se presiona el botón de finalizar la grabación hasta el momento en el que el texto aparece en la pantalla debajo de la pregunta correspondiente. Luego de esto se obtiene la media general y se analiza si existe una relación entre la cantidad de palabras y el tiempo de respuesta.

La segunda prueba consta de una comparativa entre la efectividad de transcripción entre el habla natural y el habla textual. Para realizar esta comparativa, se solicita a los sujetos que realicen la entrevista médica con el algoritmo dos veces. La primera vez lo hacen respondiendo a las preguntas únicamente recordando, esto es el habla natural ya que no tienen ningún texto preparado del cual leer entonces hay presencia de muletillas, errores gramaticales y demás. Para la segunda ocasión, se les pide que escriban ya con la forma de expresión correcta sus respuestas y vuelvan a realizar la entrevista leyendo directamente del texto sus respuestas, lo cual es habla textual. Se compara la efectividad del algoritmo para la transcripción de la entrevista completa en ambos casos para determinar si existe una variación cuando la expresión verbal no tiene un ritmo constante y pronunciación perfecta.

7.3. Algoritmo de clasificación

7.3.1. Construcción del algoritmo

La construcción del algoritmo de clasificación se divide en dos partes, el entrenamiento del algoritmo y la validación o prueba. Primero se implementa en el código el paquete de código abierto ML.NET (Microsoft, WA, USA) de modelos de aprendizaje automático. Luego se construye un archivo .csv con 624 entradas distintas para entrenamiento. Cada entrada correspondía a las 10 respuestas al esquema de entrevista junto al grupo (Traumatología, Cirugía y Medicina Interna) que debía ser asignado el paciente dependiendo del caso. Todas las entradas fueron generadas utilizando el modelo de inteligencia artificial GPT- 4 (OpenAI, CA, USA) en cantidad equitativa para las tres áreas médicas elegidas (Figura 14). Este archivo se carga al modelo dividiéndolo en las 11 entradas (10 preguntas y la clasificación) y se utiliza para entrenar el modelo.

```
Answer1,Answer2,Answer3,Answer4,Answer5,Answer6,Answer7,Answer8,Answer9,Answer10,Group
"Me duele la rodilla","Tuve una caída","Hace tres semanas","Subir escaleras lo empeora","Rodilla","No","No","No","No","No
fumo",Traumatología
"Me operaron de apéndice","Cirugía laparoscópica","Hace un mes","Comer lo empeora","Abdomen","No","No","No","No","No
fumo",Cirugía
"Me siento muy cansado","Tengo anemia","Hace un mes","El ejercicio lo empeora","Cuerpo","No","No","No","No","No
fumo",Medicina Interna
"Me fracturé el brazo","Accidente en bicicleta","Hace dos semanas","Moverlo lo empeora","Brazo","No","No","No","No","No
fumo",Traumatología
"Me operaron de hernia inguinal","Cirugía abierta","Hace un mes","Levantar peso lo empeora","Abdomen","No","No","No","No","No
fumo",Cirugía
"Me siento mareado","Tengo hipertensión","Desde hace tres semanas","El estrés lo empeora","Cabeza","No","No","No","No","No
fumo",Medicina Interna
"Me duele el pie","Me torcí el tobillo","Hace dos semanas","Caminar lo empeora","Pie","No","No","No","No","No
fumo",Traumatología
"Me operaron de vesícula","Cirugía laparoscópica","Hace un mes","Comer lo empeora","Abdomen","No","No","No","No","No
fumo",Cirugía
"Me siento débil","Me diagnosticaron insuficiencia renal","Hace un mes","El esfuerzo físico lo
empeora","Cuerpo","No","No","No","No","No fumo",Medicina Interna
"Me duele el hombro","Me lesioné jugando baloncesto","Hace un mes","Levantar el brazo lo
empeora","Hombro","No","No","No","No","No fumo",Traumatología
"Me operaron de úlcera gástrica","Cirugía abierta","Hace tres semanas","Comer lo empeora","Estómago","No","No","No","No","No
fumo",Cirugía
"Me siento agotado","Me diagnosticaron diabetes","Desde hace dos meses","El esfuerzo lo
empeora","Cuerpo","No","No","No","No","No fumo",Medicina Interna
"Me fracturé la pierna","Accidente de moto","Hace dos semanas","Caminar lo empeora","Pierna","No","No","No","No","No
fumo",Traumatología
"Me operaron de hernia umbilical","Cirugía abierta","Hace un mes","Levantarse lo empeora","Abdomen","No","No","No","No","No
```

Figura 14: Conjunto de entradas en archivo .csv utilizadas para el entrenamiento del modelo de clasificación

La clasificación de los casos de entrenamiento es realizada por el modelo de IA basándose en las características de los síntomas, partes del cuerpo afectadas, la enfermedad o lesión explicada. Por ejemplo, a traumatología las lesiones físicas, accidentes, fracturas, lesiones en extremidades o articulaciones; a medicina interna, los problemas sistémicos, infecciones, enfermedad respiratoria, crónicas o generales; por último, a cirugía, intervenciones quirúrgicas pasadas o futuras, reparaciones internas o lesiones de gravedad. Para la predicción, se carga un archivo .csv de prueba con una variedad de casos, se guarda la predicción de cada caso y se compara si la predicción coincide con el grupo previamente asignado. Se cuentan los aciertos por cada grupo clasificatorio y en un mensaje se despliega el resumen de aciertos por grupo.

7.3.2. Pruebas de efectividad de clasificación

La primera prueba de efectividad, se realiza con casos generador por el mismo modelo de IA, estos casos son distintos a los que ya se utilizan para el entrenamiento del modelo, pero la forma en la que los genera es la misma. Por lo tanto, tienen una gran similitud a la forma en la que están escritas las respuestas. Estos casos se añaden al archivo de validación y se corre el programa para obtener el resumen del acierto por grupo. Se realiza este proceso cinco veces consecutivas utilizando 60 casos distintos con igual cantidad de casos por cada grupo clasificatorio para un total de 300 casos analizados. Luego de esto, se calcula el porcentaje de efectividad basado en la cantidad de aciertos por total de aciertos, tanto para cada uno de los grupos como la efectividad general de todos los casos analizados.

Luego, se realiza el mismo tipo de prueba de efectividad, pero para casos reales. En esta prueba se transforman tanto los enterados de caso realizados por estudiantes de la Universidad Rafael Landívar y la Universidad del Valle de Guatemala como vivencias personales en respuestas a la entrevista. Para esto, se utiliza el reporte de cada caso o historia y se responde en primera persona cada una de las respuestas a las preguntas del esquema de

entrevista final con la información correspondiente. Para esta prueba, se utiliza un único archivo .csv con 30 casos distintos con igual cantidad de cada grupo para medir el porcentaje de efectividad de cada uno de los grupo y la efectividad general.

7.4. Algoritmo final integrado

En el algoritmo final se integran las tres partes fundamentales de este proyecto, el esquema de entrevista, el algoritmo de reconocimiento de voz y el de clasificación. Para esto, primero se configuran los botones dentro de la interfaz gráfica del algoritmo de reconocimiento y transcripción de voz para que la entrevista siga un orden cronológico y aparezcan en pantalla tanto las preguntas como las instrucciones de como llevar a cabo la entrevista de manera personal por parte del paciente. Este tipo de instrucciones indican que botones presionar en cada momento, cuando proseguir con las demás preguntas y como finalizar la entrevista para generar el reporte. Luego, se introduce el código perteneciente al algoritmo de clasificación y se configura el botón de finalización de entrevista para extraer todas las respuestas de la entrevista completa para ingresarla en el modelo y realizar la predicción. Por último, se especifica el directorio y se crea un archivo PDF donde se concatena tanto la entrevista completa con sus preguntas y respuestas como la sugerencia de clasificación.

CAPÍTULO 8

Resultados

8.1. Encuestas a personal médico

De las encuestas se obtuvieron una serie de preguntas que el personal médico considero relevante consultar al momento de llevar a cabo una entrevista preliminar para poder determinar la especialidad a la cual se debe de clasificar el paciente. Se recopilaron tanto estas preguntas acompañadas con la cantidad de veces que se repitieron en todo el conjunto de respuestas como el mapeo por color del tipo de pregunta dentro de las diez preguntas que escogió cada uno de los individuos encuestados (Figura 15).

Los tipos de preguntas que más se repitieron dentro de las respuestas fueron: preguntas para determinar cuales son los principales síntomas que experimenta y preguntas específicas de alguna parte del cuerpo para situar con mayor precisión el problema (11 veces) y preguntas sobre antecedentes, inicio de síntomas o momento del accidente (9 veces). De igual forma, dos preguntas solo aparecieron una vez pero se les asignó una categoría ya que pueden tener relevancia en la clasificación de un paciente como lo son consultar si existe dolor y en caso sea mujer si se encuentra embarazada.

Además de estas, 16 de las preguntas presentes en la encuesta no se repitieron y fueron consideras como aisladas. Dentro de los resultados es posible identificar que existen algunos tipos de preguntas que se repitieron más de una vez para un mismo individuo, estas fueron las preguntas relacionadas a partes del cuerpo, síntomas, operaciones o localización del dolor, en muchas ocasiones preguntándolo de una forma diferente o de forma complementaria a la pregunta realizada con anterioridad. Al tomar todo esto en cuenta se determinó el esquema final de la entrevista (Figura 16) que fue utilizado en la versión final integrada con los algoritmos de reconocimiento de voz y clasificación del paciente.

Tipo de Pregunta	Conteo
¿Cuáles son los síntomas que lo traen a consulta?	11
¿Qué le sucedió o cual es el motivo de su consulta?	8
Consulta sobre la edad o el sexo del paciente	5
¿Cuándo iniciaron los síntomas?	9
Preguntas específicas de alguna parte del cuerpo	11
Consultar por antecedentes personales y familiares	9
¿Qué medicametos consumió o consume regularmente?	8
¿Sufrió algún tipo de accidente?	9
¿Padece de alguna enfermedad?	7
Preguntas aisladas o de única respuesta	16
¿Se encuentra usted embarazada? Aplica únicamente a mujeres	1
¿Cómo es el dolor?	5
¿Qué intensifica la molestia o dolor?	5
Consultas sobre operaciones pasadas o futuras	5
¿Dónde se encuentra ubicado el dolor?	4
¿Presenta dolor?	1
¿Algún médico ya lo vio con anterioridad?	7
¿Ha presentado síntomas asociados como vómitos, náuseas, fiebre?	4
Consulta sobre la evolución de los síntomas	2
¿Ha realizado algún examen o laboratorio complementarios?	3

	P#1	P#2	P#3	P#4	P#5	P#6	P#7	P#8	P#9	P#10
Individuo 1										
Individuo 2										
Individuo 3										
Individuo 4										
Individuo 5										
Individuo 6										
Individuo 7										
Individuo 8										
Individuo 9										
Individuo 10										
Individuo 11										
Individuo 12										
Individuo 13										

Figura 15: Tipos de preguntas elegidas por el personal de salud y su respectivo mapeo dentro de las 10 preguntas de cada individuo encuestado

1. ¿Cuáles son los síntomas que lo traen a consulta?

2. ¿Qué le sucedió o cual es el motivo de su consulta?

3. ¿Hace cuanto iniciaron los síntomas o tuvo el accidente? (cuatro días, una semana, un mes…)

4. ¿Qué intensifica la molestia o dolor?

5. ¿Dónde se encuentra ubicado el dolor o complicación? (Parte del cuerpo)

6. ¿Usted padece de alguna enfermedad crónica? Si no padece responda NO, si padece alguna enfermedad responda únicamente que enfermedades.

7. ¿Sus papás, abuelos, tíos o hermanos padecen de alguna enfermedad crónica? Si no padecen responda NO, si padecen alguna enfermedad responda únicamente que enfermedades.

8. ¿Alguna vez lo han operado quirúrgicamente? Si no lo han operado responda NO, si sí lo han operado responda únicamente de que lo han operado.

9. ¿Alguna vez lo han ha sufrido un accidente de gravedad? Si no es así responda NO, si sí lo ha sufrido responda únicamente que parte del cuerpo se accidentó.

10. ¿Fuma y/o consume bebidas alcohólicas con frecuencia?

Figura 16: Esquema final de 10 preguntas para la entrevista

8.2. Pruebas de reconocimiento de voz

Los porcentajes de efectividad en la transcripción para cada uno de los individuos en las pruebas de reconocimiento fueron agrupados por versión de algoritmo con la respectiva media del grupo representada por un asterisco rojo y la mediana representada por una línea de mayor grosor en la parte interior de la caja (Figura 17). Esto permite identificar el efecto que tiene cada una de las mejoras en la versión del algoritmo con respecto a las anteriores y si este efecto es positivo o negativo en términos de alcanzar una efectividad mayor y reducir la cantidad de errores al momento de la transcripción. Viendo la media de cada grupo desde el algoritmo básico hasta el algoritmo que utiliza inteligencia artificial para la transcripción es posible notar una tendencia al alza.

La diferencia entre las medias de los primeros tres grupos es relativamente baja, entre 3 % y 4 % de una versión a la otra. Por otro lado, la diferencia entre los siguientes dos es casi tres veces mayor (15 %) que las otras variaciones de promedio entre grupos. La última versión supera en 46 % a la mayor de todas. Esto denota como el micrófono profesional tiene un efecto significativo sobre la capacidad de transcripción del algoritmo de reconocimiento. También se puede identificar que el cambio mayor se dio al momento de cambiar el sistema con el cual se realiza la transcripción de un sistema local a un servidor en la nube que emplea mecanismos más complejos. De igual forma se puede apreciar que los resultados individuales de cada una de las personas analizadas en la mayoría de los casos mantuvieron cierta constancia con respecto a cómo se sitúa con relación a las demás personas, los individuos con resultados por encima o por debajo de la media mantuvieron estas métricas en los cinco casos.

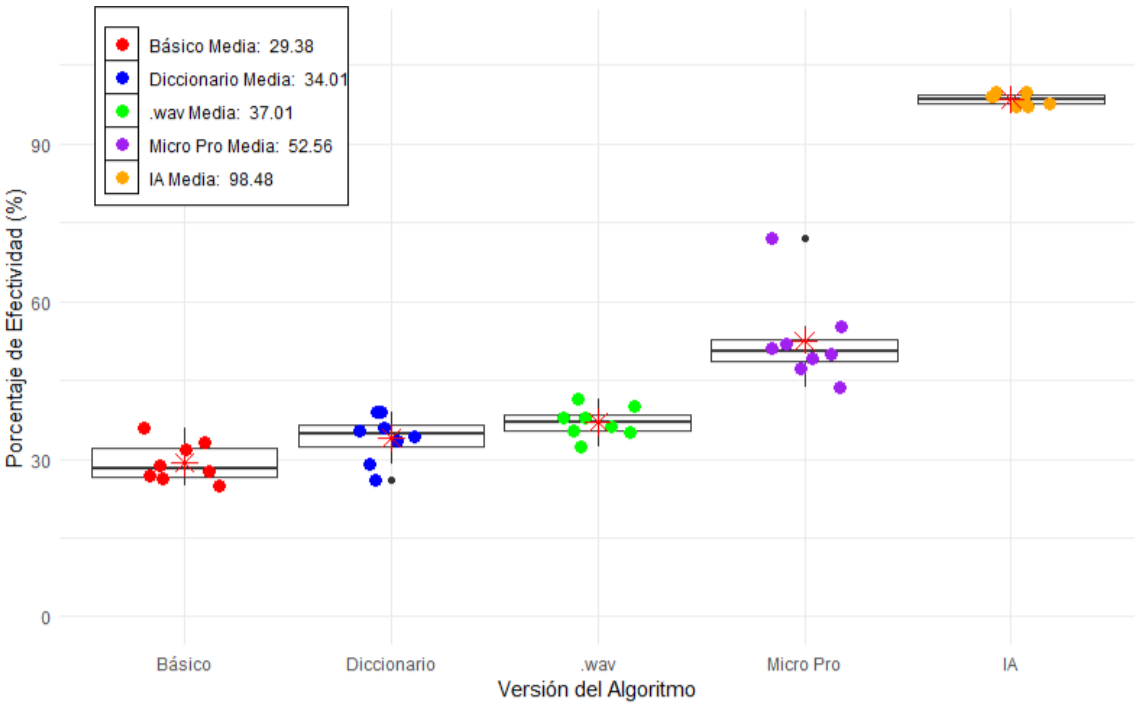


Figura 17: Gráfico de caja del porcentaje de efectividad de transcripción agrupada por versión de algoritmo

Al realizar las diferentes pruebas en cada una de las versiones los tiempos de respuesta, el tiempo que transcurría para realizar la transcripción, variaba de versión a versión. Por esta razón, se consideró de relevancia medir esta variable para el algoritmo seleccionado para la versión final. Se midió la velocidad de transcripción con relación a la cantidad de palabras a transcribir. Así es posible identificar si la cantidad de palabras tiene un efecto directo sobre el tiempo. La importancia de medir esta variable recae en que cuando se implementa este algoritmo en un escenario real es importante saber cuanto tiempo tarda en reaccionar el algoritmo cada vez que se responde una pregunta.

Al comparar los resultados de la velocidad por cantidad de palabras (Figura 18), no es posible identificar un patrón claro al aumentar la cantidad de palabras. No presenta ningún tipo de tendencia ya sea al alza o a la baja. La media más alta identificada dentro del grupo se encontró para el caso de 6 palabras que es una muestra intermedia entre los seis tipos evaluados. Por el contrario, la muestra con la menor media se dio para el caso de 18 palabras que es la mayor cantidad de palabras que fue evaluada para este experimento. Al evaluar las medias de todos los casos,representadas por los puntos de color en cada una de las cajas, si es posible identificar un patrón con relación a que independientemente de cual sea la cantidad de palabras transcritas el tiempo de respuesta es muy similar, casi todas se encuentran alrededor del mismo rango. Por lo tanto, fue posible determinar la media general de todos los casos la cual al encontrarse todas en un rango bastante similar es bastante representativa del tiempo que suele tomarse el algoritmo en transcribir un texto, esta media, representada por la línea roja punteada, tuvo un valor de 17.6 segundos.

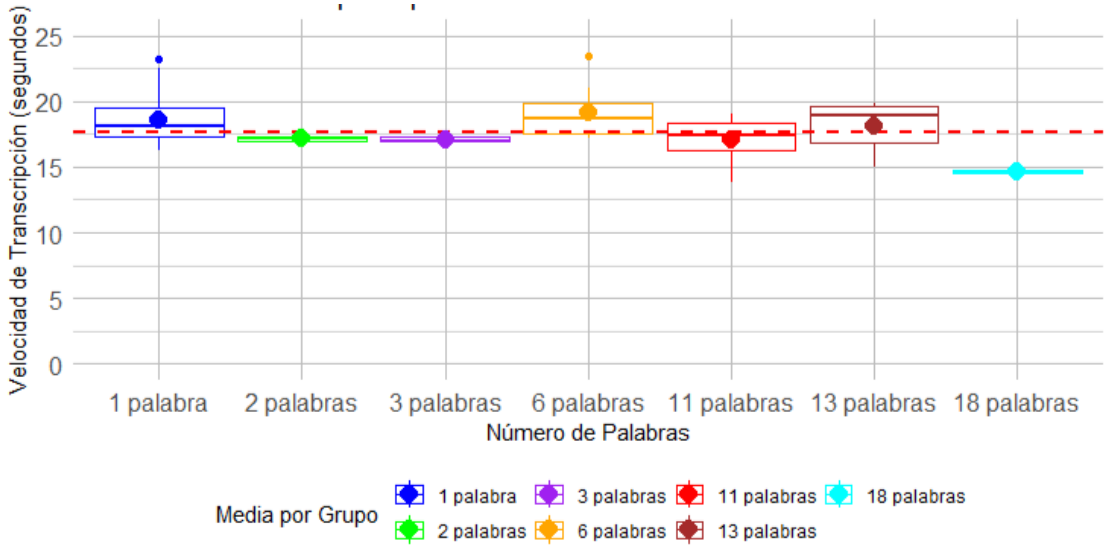


Figura 18: Gráfico de caja de la velocidad de transcripción por número de palabras medida en segundos

La transcripción puede variar dependiendo de diferentes factores al momento de que se obtiene la señal de audio. Cuando el proceso de reconocimiento tiene variaciones en esa misma entrada esto se puede llegar a ver reflejado en la capacidad o efectividad de transcripción de un modelo. Ejemplo de esto es cuando se compara la efectividad que puede tener un sistema de reconocimiento y transcripción utilizando habla natural o habla textual. La comparativa de estos dos tipos de entrada si presenta cierta diferencia (Figura 19), tanto la variación que existe entre mediciones de la efectividad la cual se puede apreciar en el tamaño de caja de cada tipo como la media. No obstante, ambas medias son bastante altas, 99.9 % para el habla textual y 99.34 % para el habla natural, prácticamente con un porcentaje de efectividad casi ideal. Comparando entre ambos tipos la variación es prácticamente nula, una diferencia de 0.54 % es relativamente baja tomando en cuenta ambos tipos de entrada.

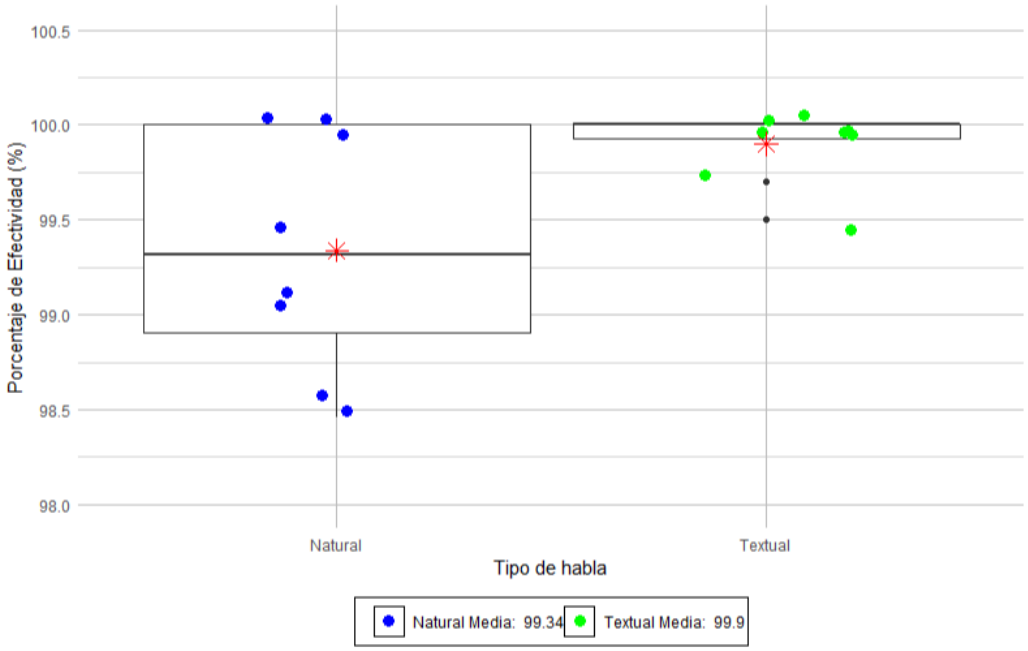


Figura 19: Gráfico de caja comparativo entre la efectividad de transcripción del habla natural y textual

8.3. Pruebas a algoritmo de clasificación

Para medir la efectividad de un modelo se necesita saber como realiza la clasificación en términos de efectividad, ya sea entre todos los grupos como para cada uno. Los porcentajes de efectividad de clasificación del modelo utilizando entradas generadas de la misma manera que los casos de entrenamiento (Figura 20) se dividió para cada uno de los grupos/especialidad médica (Traumatología, Medicina Interna, Cirugía) así como también la efectividad general de todos los aciertos sin diferenciar entre grupos. Con esto es posible identificar si existe una diferencia entre cada una de las especialidades, que tan efectiva es la clasificación y en términos generales en donde se sitúa este porcentaje y su media la cual se encuentra identificada como el punto negro perteneciente a cada caja o grupo.

Entre los tres la especialidad con los mejores resultados fue la de Medicina Interna que obtuvo un porcentaje de efectividad del 100 %. El segundo de los tres fue el de cirugía donde la mayoría de los datos tuvieron una efectividad total, pero presentó un dato atípico alrededor del 88 % el cual generó una media de 97.6 %. El último fue el de Traumatología donde los resultados si sufrieron una desviación mucho mayor con una media de 90.4 % que si tiene una diferencia considerables con relación a los otros dos grupos. Esto generó una media general de 96 % siendo esto un valor bastante elevado de efectividad para la muestra utilizada. En estos resultados es posible apreciar tres casos distintos de distribución ya que en el primero todos los datos tuvieron una efectividad ideal. Para los casos de cirugía podría parecer que es ideal pero el dato atípico tan desviado de la media no coincide con la tendencia. Por último, los casos de traumatología si se encuentran distribuidos uniformemente alrededor de todo el rango expuesto.

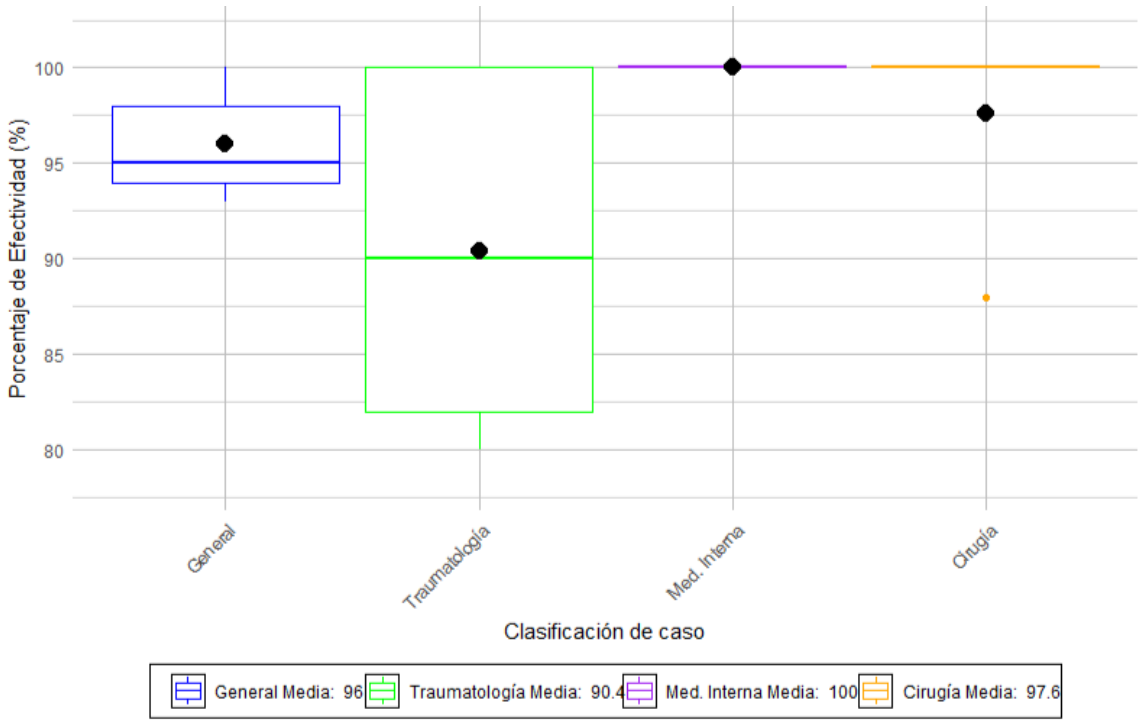


Figura 20: Gráfico de caja del porcentaje de efectividad de clasificación por grupo o especialidad médica evaluada para los casos generados de la misma manera que los casos de entrenamiento

Al evaluar la efectividad de clasificación para los casos reales se alteró la forma en la que se presentan los datos con relación a la misma medición de efectividad en los casos generados de igual forma que los de entrenamiento. Para los casos reales extraídos de enterados de casos médicos se genera un gráfico de barras (Figura 21) con una muestra menor que es representativa de todos los casos clasificados. Es posible apreciar que para estos casos la tendencia del orden de efectividad por especialidad sigue siendo el mismo. El mayor de todos es el de Medicina Interna con una efectividad del 60 %, luego cirugía con un 50 % y el último de todos es Traumatología con un 40 % de efectividad. Dando como resultado un 50 % de efectividad general para la evaluación de la capacidad de clasificación del modelo con casos reales. Si se compara la media obtenida para los casos similares a los de entrenamiento (Figura 20) con estos resultados es posible apreciar una disminución media del 46 %. Este valor de disminución es bastante alto ya que representa que en las tres especialidades disminuyeron casi a la mitad la efectividad de clasificación cuando se evalúa el modelo con casos reales.

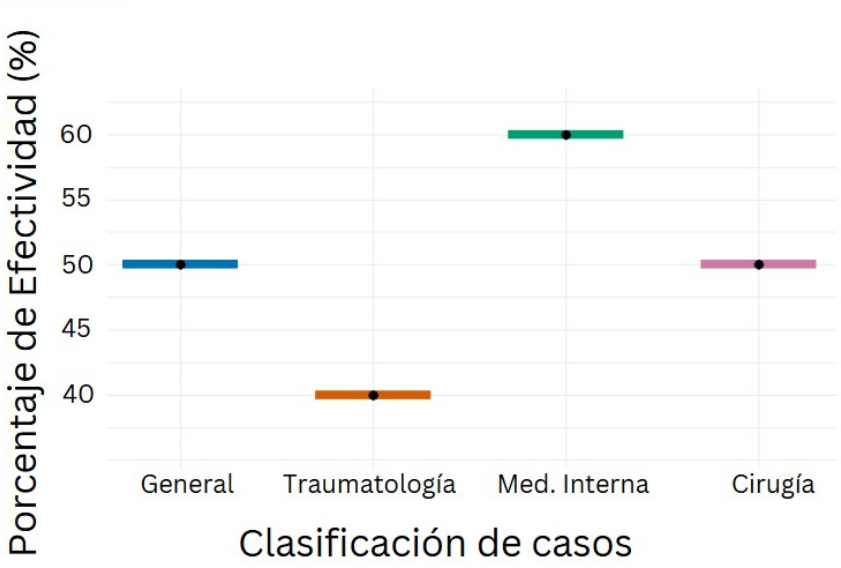


Figura 21: Gráfico de caja del porcentaje de efectividad de clasificación por grupo o especialidad médica evaluada para los casos reales

Con base en la integración de ambos algoritmos fue posible conseguir como resultado un algoritmo como versión final que realiza tanto los procesos de reconocimiento y transcripción de la voz como la clasificación. Cuando finaliza la entrevista se presiona el botón de salida del formulario y el modelo de clasificación utiliza las respuestas como entradas y genera una única predicción. También se genera un reporte (Figura 22) con la información transcrita y la sugerencia de clasificación que representa el final de la entrevista.

1.¿Cuáles son los síntomas que lo traen a consulta?

Tengo dolor en las piernas, sobre todo en la derecha

2.¿Qué le sucedió o cual es el motivo de su consulta?

El dolor me empezó de la nada y no ha parado

3.¿Hace cuanto iniciaron los síntomas o tuvo el accidente? (8 horas, 4 días, 1 mes…)

24 horas

4.¿Qué intensifica la molestia o dolor?

Caminar mucho tiempo

5.¿Dónde se encuentra ubicado el dolor o complicación? (Parte del cuerpo)
Ambas piernas

6.¿Usted padece de alguna enfermedad crónica? Si no padece responda NO,
si padece alguna enfermedad responda únicamente que enfermedades.

Diabetes, EPOC, fibrosis pulmonar

7.¿Sus papás, abuelos, tíos o hermanos padecen de alguna enfermedad crónica? Si no padecen responda NO, si padecen alguna enfermedad responda únicamente que enfermedades.

No

8.¿Alguna vez lo han operado quirúrgicamente? Si no lo han operado responda NO,
si sí lo han operado responda únicamente de que lo han operado.

Resección transuretral

9.¿Alguna vez ha sufrido un accidente de gravedad? Si no es así responda NO, si sí lo ha sufrido responda únicamente que parte del cuerpo se accidentó.

No

10.¿Fuma y/o consume bebidas alcohólicas con frecuencia?

No

Medicina Interna

Figura 22: Contenido del reporte final en PDF con la información transcrita y la sugerencia de clasificación al final

CAPÍTULO 9

Discusión de Resultados

9.1. Esquema de la entrevista

Para la determinación de las preguntas tuvieron una gran relevancia los resultados obtenidos a partir de la encuesta sobre la cantidad de veces que se repitió cada tipo de pregunta (Figura 15). La importancia que le dan los médicos a cada pregunta se ve reflejada en esta estadística y su relación con la clasificación se ve determinada por diferentes factores. Una de las preguntas más repetidas fue consultar sobre las partes del cuerpo afectadas. La mayor distinción la genera en diferenciar los casos de Traumatología de los demás ya que estos se centran en lesiones de huesos, articulaciones, tendones y músculos, casos que no tratan las otras dos especialidades [54]. La Cirugía y Medicina Interna se enfocan en sistemas internos del cuerpo, pero dependiendo del tipo problema o enfermedad que se presente se realiza la distinción entre ambas áreas [55]. Esto refleja la razón por la cual este tipo de pregunta es considerada como una de las principales para realizar clasificación por parte del personal médico.

Ahora con relación a la pregunta de cuales son los síntomas que lo traen a consulta, esta pregunta si permite realizar distinciones entre las tres especialidades ya que los síntomas de cada una presentan diferencias. A Traumatología se refieren los casos relacionados a daño a nivel muscular o esquelético, dolor intenso en estas zonas. La presencia de masas anormales, disfunción en órganos o dolor severo son síntomas relacionados a casos de cirugía. Por otro lado, fallas multisistémicas, dolor crónico, fatiga o fiebre son algunos de los síntomas tratados por medicina interna. La diferencia entre este tipo de síntomas hace que esta pregunta sea un buen indicador de cual podría ser la clasificación adecuada de cada caso [56] [57].

9.2. Algoritmo de reconocimiento de voz

Los algoritmos de reconocimiento de voz varían su efectividad de transcripción (Figura 17) con base en que tan específicos sean o las herramientas que utilicen para facilitar el reconocimiento. La mejora de la primera versión a la segunda está relacionada a que un reconocimiento enfocado o específico reduce el tiempo de procesamiento y la tasa de error por palabra significativamente, mientras más enfocado sea mayor es la efectividad de reconocimiento [58]. Para la tercera versión la diferencia la marca el filtro pasabandas, este tipo de filtro generó que el rango de frecuencias analizado se reduzca. Esto elimina una gran parte del ruido de fondo y mantiene las características más importantes de la voz. Por ende, se obtiene una efectividad de transcripción mayor al contar con una señal mucho más limpia [59].

Al comparar las primeras versiones con la que utiliza un micrófono profesional, se notó una mejora considerable en la efectividad de transcripción. Esto se debió a que desde la captura de la señal la sensibilidad de captación, la eliminación del ruido y la supresión de reverberaciones están presentes aislando de una mejor manera la voz [60]. Esto sumado al posterior procesamiento con filtrado hace que la transcripción sea mucho más precisa. Para la versión elegida que fue la que emplea IA la mejora en la efectividad fue abismal, muy por encima de la media de todas las demás versiones, esto se debió a una gran cantidad de factores. Los servidores de IA tienen acceso a muchos más datos que un motor de reconocimiento local como el de las demás versiones. Utilizan técnicas de procesamiento paralelo que aumenta su capacidad, están en procesos constantes de auto aprendizaje para mejorar su efectividad. Tienen acceso a algoritmos de cancelación de ruido o aprendizaje de patrones mucho más complejos que los motores locales por que estos últimos tienen limitaciones de almacenamiento y *hardware* [61].

Verificando la velocidad de transcripción (Figura 18), se pudo identificar un tiempo bastante similar sin importar la cantidad de palabras que se transcribieron. En realidad, la cantidad de palabras si tiene un efecto sobre el tiempo que tarda en procesarse la transcripción, pero es significativa cuando se comparan textos de longitudes mucho más grandes. En este caso donde las frases o respuestas eran de máximo 18 palabras no va a existir una diferencia marcada porque este tipo de algoritmos son lo suficientemente optimizados para procesar por igual entradas así de cortas. Por esa razón no se identificó ningún aumento en el tiempo de procesamiento del algoritmo [62]. Tomando en consideración que el archivo se sube al servidor, se transcribe y luego regresa el texto transcrito al algoritmo local el tiempo promedio obtenido para este proyecto es aceptable ya que todo este proceso suele tomar entre 10 y 20 segundos [63], para este proyecto la media es de 17.6 segundos. No obstante, es recomendable contar con tiempos menores para que la experiencia de usuario sea más agradable.

La efectividad de transcripción del habla natural fue levemente menor a la del habla textual, (Figura 19) la diferencia es poco significativa, pero es un factor a tomar en cuenta que en una entrevista normal la efectividad pueda reducirse. Los modelos de reconocimiento de voz entrenados con datos y audios de habla con buena pronunciación y ritmo constante tienden a presentar mejores resultados de efectividad con habla textual, la mayoría de los modelos son entrenados con este tipo de datos. La tasa de error por palabra es mayor para el habla natural por la presencia de interrupciones, poca imprevisibilidad, ritmo desajustado,

mala pronunciación o acentos [64]. Para el habla natural el porcentaje de error fue menor al 1 % y la mayoría de sistemas de reconocimiento automático reconocen porcentajes de error menores al 10 % como aceptables bajo condiciones ideales de habla natural [65].

9.3. Algoritmo de clasificación

Los porcentajes de efectividad al probar el algoritmo de clasificación con datos similares a los utilizados en su entrenamiento fueron bastante altos, con medias por encima del 90 % de efectividad (Figura 20). Estos resultados reflejan una gran capacidad de predicción y clasificación del algoritmo cuando se enfrenta a casos o respuestas de entrevistas que tienen un gran parecido a los datos ya aprendidos. Demuestra que el algoritmo es capaz de determinar con base en su entrenamiento la clasificación que mejor se ajuste al caso que analiza. No obstante, a pesar de que esto es un buen indicador de su capacidad de clasificación esto es bastante común en la mayoría de los algoritmos de este tipo. Esto en muchos casos se puede deber a un proceso de *overfitting* que sucede cuando el modelo memoriza la forma de los datos en vez de aprender a generalizar a partir de estos datos, produciendo así una ilusión de una alta efectividad [66].

La caída de efectividad fue drástica al probar el modelo con datos de casos reales que tienen una diferencia significativa con las entradas a las que está acostumbrado el modelo (Figura 21). Los porcentajes de efectividad de las diferentes especialidades analizadas disminuyeron en un 46 % de media. Este tipo de modelos suelen ser bastante sensibles a datos que se alejan de lo esperado, la robustez del modelo va a aumentar su nivel de adaptabilidad a este tipo de escenarios. Los modelos aprenden patrones específicos que extraen de su entrenamiento entonces casos ajenos a esos patrones son difíciles de clasificar. La caída de efectividad es muy variable, pero puede oscilar entre un 10 % y un 50 %. La caída presente en este proyecto entra dentro de este rango en su extremo superior lo cual puede considerarse aceptable o normal, pero demuestra que existe una falta de diversidad y representatividad en los datos de aprendizaje [67].

Los modelos de clasificación tienen como característica principal su efectividad de clasificación. Se miden con base en este parámetro y cobra aún mayor importancia cuando hablamos de modelos que son empleados en un ambiente clínico. En los resultados descritos el porcentaje de efectividad general fue del 50 %, este valor puede considerarse bajo ya que de cada dos casos clasifica de manera correcta únicamente uno. Se sugiere que para modelos utilizados en un entorno clínico se debe de alcanzar un 80 % de precisión en la clasificación, valor que se encuentra bastante por encima del obtenido en este proyecto. Esto refleja la necesidad de un modelo mucho más efectivo, que sea capaz de predecir con mayor precisión casos reales. Por el tipo de importancia que se le da a la clasificación de un paciente la cual no tiene repercusiones significativas en caso de error un 80 % es aceptable, pero si se desea dar más peso a esta decisión el valor puede aumentar hasta 85 % [68].

CAPÍTULO 10

Conclusiones

El esquema de la entrevista se obtuvo con base en el conocimiento y opinión de un grupo de médicos mayor al estipulado dentro de los objetivos, logrando así construir una entrevista con fundamento y las bases necesarias para contribuir positivamente a la identificación del problema, los síntomas y la clasificación del paciente. Esto se logró por medio de diferentes fuentes de recaudación de información como entrevistas y encuestas realizadas al mismo personal médico. El esquema conseguido tiene repercusiones directas sobre la efectividad de clasificación del algoritmo correspondiente dentro de este proyecto. Esto debido a que las respuestas son las entradas del algoritmo que se utiliza para determinar la clasificación.

Únicamente se evaluó la influencia de dos métodos de grabación en la calidad del audio. Se debió a que, al momento de analizar la efectividad producida por estos dos métodos, siendo uno básico con el micrófono de la computadora y el otro uno profesional, la diferencia de efectividad no se consideró lo suficientemente significativa como para considerar realizar una prueba más con un método intermedio. Esto implicaba realizar una nueva serie de pruebas para todo el grupo de prueba cuando los resultados seguramente reflejarían un leve aumento en la capacidad de transcripción. Por lo tanto, se tomó la decisión de únicamente evaluar un método básico y uno profesional e identificar la diferencia porcentual de efectividad entre ambos métodos. El hallazgo de la poca influencia que tenía ese cambio de método impulsó la creación de la versión que utiliza inteligencia artificial para realizar la transcripción del audio.

Fue posible determinar la versión de algoritmo de reconocimiento de voz más adecuado para la versión final. A lo largo de la realización del proyecto, se evaluó la efectividad de transcripción de cinco versiones distintas para realizar una comparativa y así decidir cual de las estas era la indicada. La elegida fue la que utiliza servidores en la nube con motores impulsados por inteligencia artificial, esta fue la que presentó la efectividad de transcripción mayor entre todas las versiones; contando esta con una diferencia de 46 % de efectividad por encima de la media alcanzada por la segunda versión con mayor efectividad. Un tiempo de respuesta dentro de los rangos aceptables para este tipo de procesamiento y una reducción mínima del porcentaje de efectividad medio al evaluarlo con habla natural como la que se

aprecia en una entrevista. Este algoritmo cumple con los requerimientos suficientes para ser aceptado en su utilización como motor de transcripción. No obstante, por el tipo de algoritmo seleccionado, la inclusión de librerías y diccionarios de fuente abierta en español no fue implementada directamente si no que viene implícita en el servidor de AssemblyAI empleado.

Al finalizar el proyecto, fue posible integrar de manera exitosa las tres bases que lo conforman, el esquema de entrevista, el algoritmo de reconocimiento de voz y el algoritmo de clasificación de pacientes. Se evaluó exitosamente este algoritmo final con base en su capacidad de clasificación tanto para casos similares a los de entrenamiento como para casos reales extraídos de enterados de casos médicos o experiencias de los mismos usuarios de prueba. Disminuyendo considerablemente su capacidad de clasificación para casos reales. Esta disminución no hace recomendable la utilización del modelo de clasificación para usos clínicos. A partir de esta integración se logró generar un reporte final donde se encuentran todas las preguntas y respuestas de la entrevista, dentro de estas los síntomas. El reporte incluye también el posible especialista a visitar, este último derivado del procedimiento realizado por el algoritmo de clasificación.

Desarrollar el algoritmo de realización y análisis de entrevistas utilizando reconocimiento de voz para transcribir todas las respuestas con base en el esquema de la entrevista y un algoritmo de clasificación para realizar una sugerencia de especialidad médica en base a esas mismas respuestas fue un gran reto. Los resultados obtenidos permiten inferir que la efectividad de transcripción a la hora del reconocimiento de voz es más que adecuada para este tipo de implementación. Sin embargo, analizando la clasificación de pacientes, la versión final evidencia poder clasificar satisfactoriamente casos similares a los de su entrenamiento, pero cuenta con una gran área de mejora al evaluarlo con casos de la vida real que son a los cuales se enfrentaría en este tipo de aplicación. Este proyecto es una buena base para la futura creación de un algoritmo de uso clínico que revolucione la forma en la que actualmente se realiza este tipo de procedimientos.

CAPÍTULO 11

Recomendaciones

Para determinar el mejor esquema posible de las preguntas sería recomendable realizar más análisis cualitativos y cuantitativos a las preguntas obtenidas ya que es una forma en la que se puede verificar que tan útiles son las preguntas para este tipo de entrevista. Una de las evaluaciones que mayor impacto podrían tener sobre la construcción del esquema sería hacer varios esquemas con diferentes combinaciones de preguntas y analizar en el algoritmo de clasificación cual es el esquema que genera los mejores resultados. Contar con una buena muestra para encuestas y pruebas es beneficioso para dotar de confiabilidad a los resultados obtenidos, se recomienda comprobar la veracidad de los resultados obtenidos para este proyecto realizando el mismo tipo de pruebas y encuestas con una muestra mayor.

Otro tipo de encuestas que útil para sustentar la justificación de este proyecto podrían ser encuestas sobre el nivel de satisfacción. Esta se realizaría tanto a población general para analizar la disposición a cambiar del método tradicional de conducir entrevistas a uno computacional y si confiarían en sus resultados. La segunda sería al personal médico, para estudiar su adecuación a este método y que tan beneficioso es en términos de liberar a su personal para poder enfocarse en otros asuntos. Con relación al motor de reconocimiento el detalle que podría mejorarse del sistema actual sería la utilización de un motor local en vez de un servidor, ya fuese que utilice o no inteligencia artificial, pero al usar uno local el tiempo de transcripción se reduce considerablemente haciendo al sistema más eficiente.

La parte con mayor área de mejora es la del algoritmo de clasificación. Para este caso por cuestiones de la cantidad de datos que necesita un modelo de este tipo para generar predicciones acertadas se utilizaron casos generados con inteligencia artificial ya que esto permitía crear altas cantidades de casos diferentes de una manera veloz. Para este proyecto que su meta es trabajar con casos de la vida real lo más recomendable es que el entrenamiento del modelo se haga con este tipo de datos. También lo mejor es que cuente con una mayor cantidad de entradas para su entrenamiento que las que actualmente utiliza, esto produciría un modelo mucho más robusto y con una efectividad aceptable. También se debe de crear un modelo más complejo que esté en constante aprendizaje, que cada vez que se prueba con nuevos casos vaya aprendiendo de ellos y así mejore con el paso del tiempo.

Por último, existen algunas metas que podrían estipularse para trabajar en este algoritmo que lo convertirían en una herramienta de gran utilidad y muy completo. La primera de ellas es aumentar la cantidad de especialidades de las cuales es capaz de generar una clasificación. Solo se trabajó con las tres áreas básicas que se manejan en los hospitales, pero se podría aumentar cada vez más hasta conseguir un modelo que pueda ser utilizado previo o en el lugar para consulta externa donde la cantidad de especialidades es mucho más elevada. Obviamente, esto necesita de mucha más información que ayude a que esto sea certero ya que mientras más especialidades la posibilidad de clasificar erróneamente aumenta también. La otra meta sería que, con base en información sobre distintas enfermedades, qué síntomas presentan, las características que las distinguen, etc., sea posible determinar una lista de posibles enfermedades que podrían ser el problema añadido a la clasificación del paciente. Contar con toda esta información le permite al doctor tratante tener un panorama bastante completo del caso.

CAPÍTULO 12

Bibliografía

[1] P. Spachos, S. Gregori y M. J. Deen, "Voice activated IoT devices for healthcare: Design challenges and emerging applications," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, n.º 7, págs. 3101-3107, 2022.

[2] M. A. Fox, C. J. Aschkenasi y A. Kalyanpur, "Voice recognition is here comma like it or not period," *Indian Journal of Radiology and Imaging*, vol. 23, n.º 03, págs. 191-194, 2013.

[3] A. Sattar y M. Hafeez, "The Adjunct of Voice Recognition to Medical Transcriptionist in Asian Countries–The Pros and Cons," *American Journal of Internal Medicine*, vol. 7, n.º 6, págs. 147-150, 2019.

[4] J. Zhang, J. Wu, Y. Qiu et al., "Intelligent speech technologies for transcription, disease diagnosis, and medical equipment interactive control in smart hospitals: A review," *Computers in Biology and Medicine*, vol. 153, pág. 106 517, 2023.

[5] D. M. M. Pacis, E. D. Subido y N. T. Bugtai, "Trends in telemedicine utilizing artificial intelligence," en *AIP conference proceedings*, AIP Publishing, vol. 1933, 2018.

[6] P. Baudier, C. Ammi y G. Kondrateva, "The acceptability of telemedicine cabins by the students," *Journal of innovation economics & management*, págs. I75-21, 2020.

[7] M. Wojtara, E. Rana, T. Rahman, P. Khanna y H. Singh, "Artificial intelligence in rare disease diagnosis and treatment," *Clinical and Translational Science*, vol. 16, n.º 11, págs. 2106-2111, 2023.

[8] USAID, *Guatemala Analisis del Sistema de Salud 2015*, https://2017-2020.usaid.gov/sites/default/files/documents/1862/Guatemala-Analisis-del-Sector-Publico-Salud-Esp-INFORME-COMPLETO-FINAL-Abr2016.pdf, [Accessed 17-07-2024], 2016.

[9] J. M. McMillin, "Clinical methods: the history, physical, and laboratory examinations," en *Blood glucose*, 141, Butterworth, 2024.

[10] P. Haidet y D. A. Paterniti, "Building a history rather than taking one: A perspective on information sharing during the medical interview," *Archives of internal medicine*, vol. 163, n.º 10, págs. 1134-1140, 2003.

[11] R. Mendel, E. Traut-Mattausch, E. Jonas et al., "Confirmation bias: why psychiatrists stick to wrong preliminary diagnoses," *Psychological medicine*, vol. 41, n.º 12, págs. 2651-2659, 2011.

[12] R. Featherston, L. E. Downie, A. P. Vogel y K. L. Galvin, "Decision making biases in the allied health professions: a systematic scoping review," *PLoS One*, vol. 15, n.º 10, e0240716, 2020.

[13] L. A. Siminoff, H. L. Rogers, M. D. Thomson, L. Dumenci y S. Harris-Haywood, "Doctor, what's wrong with me? Factors that delay the diagnosis of colorectal cancer," *Patient education and counseling*, vol. 84, n.º 3, págs. 352-358, 2011.

[14] A. Pradhan, K. Mehta y L. Findlater, ".ªccessibility Came by AccidentÜse of Voice-Controlled Intelligent Personal Assistants by People with Disabilities," en *Proceedings of the 2018 CHI Conference on human factors in computing systems*, 2018, págs. 1-13.

[15] C. Bokhove y C. Downey, "Automated generation of 'good enough'transcripts as a first step to transcription of audio-recorded data," *Methodological innovations*, vol. 11, n.º 2, pág. 2 059 799 118 790 743, 2018.

[16] R. Rutakumwa, J. O. Mugisha, S. Bernays et al., "Conducting in-depth interviews with and without voice recorders: a comparative analysis," *Qualitative Research*, vol. 20, n.º 5, págs. 565-581, 2020.

[17] D. McIntyre y C. K. Chow, "Waiting time as an indicator for health services under strain: a narrative review," *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, vol. 57, pág. 0 046 958 020 910 305, 2020.

[18] C. P. West, L. N. Dyrbye y T. D. Shanafelt, "Physician burnout: contributors, consequences and solutions," *Journal of internal medicine*, vol. 283, n.º 6, págs. 516-529, 2018.

[19] S. W. Yates, "Physician stress and burnout," *The American journal of medicine*, vol. 133, n.º 2, págs. 160-164, 2020.

[20] *Qué es Voz: Diccionario Médico - Clínica U. Navarra — cun.es*, https://www.cun.es/diccionario-medico/terminos/voz, [Accessed 28-07-2024].

[21] C. Sapienza y B. Hoffman, *Voice disorders*. Plural Publishing, 2020.

[22] J. Macias-Guarasa, *El sistema de producción de habla*, ene. de 2015.

[23] E. Etecé, *Aparato fonador*, https://humanidades.com/aparato-fonador/, [Accessed 06-11-2024], 2024.

[24] P. Photinos, *The Physics of Sound Waves: Music, instruments, and sound equipment*. IOP Publishing, 2021.

[25] V. V. Kadam y R. Nayak, "Basics of acoustic science," *Acoustic Textiles*, págs. 33-42, 2016.

[26] *Principle of superposition | Definition, Examples, & Facts — britannica.com*, https://www.britannica.com/science/principle-of-superposition-wave-motion, [Accessed 29-07-2024].

[27] U. of Colorado Boulder, *Creación de ondas con Fourier*, https://phet.colorado.edu/es/simulations/fourier-making-waves, [Accessed 06-11-2024], 2024.

[28] N. Kazanina, J. S. Bowers y W. Idsardi, "Phonemes: Lexical access and beyond," *Psychonomic bulletin & review*, vol. 25, n.º 2, págs. 560-585, 2018.

[29] L. G. Duncan, "Language and reading: The role of morpheme and phoneme awareness," *Current developmental disorders reports*, vol. 5, págs. 226-234, 2018.

[30] J. Trelles, *Correspondencia Entre Fonemas Y Grafemas*, https://pdfcoffee.com/correspondencia-entre-fonemas-y-grafemas--pdf-free.html, [Accessed 06-11-2024], 2024.

[31] Gemma, *CLASIFICACIÓN DE LAS VOCES HUMANAS - GemmaPedros — gemmapedros.com*, https://www.gemmapedros.com/2020/05/clasificacion-de-las-voces-humanas/, [Accessed 29-07-2024].

[32] Suzuki, *Las 6 tesituras de la voz*, https://www.suzukitalenteducation.com.mx/blog?id=46, [Accessed 06-11-2024], 2022.

[33] M. M. Uddin, N. Huynh, J. M. Vidal, K. M. Taaffe, L. D. Fredendall y J. S. Greenstein, "Evaluation of Google's voice recognition and sentence classification for health care applications," *Engineering Management Journal*, vol. 27, n.º 3, págs. 152-162, 2015.

[34] M. Dong, L. Peng, Q. Nie y W. Li, "Speech Signal Processing of Industrial Speech Recognition," en *Journal of Physics: Conference Series*, IOP Publishing, vol. 2508, 2023, pág. 012 039.

[35] Xnomind, *Teorema de Nyquist con explicación Sencilla - Teorema — teorema.top*, https://www.teorema.top/teorema-de-nyquist/, [Accessed 03-08-2024].

[36] *monolithicpower.com*, https://www.monolithicpower.com/en/learning/mpscholar/analog-to-digital-converters/introduction-to-adcs/fundamental-concepts, [Accessed 29-07-2024].

[37] *3.4. Windowing &x2014; Introduction to Speech Processing — speechprocessingbook.aalto.fi*, https://speechprocessingbook.aalto.fi/Representations/Windowing.html, [Accessed 29-07-2024].

[38] D. Jurafsky y J. H. Martin, "Automatic Speech Recognition and Text-to-Speech," n.º 30/07/2024, págs. 1-30, feb. de 2024. dirección: https://web.stanford.edu/~jurafsky/slp3/16.pdf.

[39] B. K, *Hamming window and frequency response*, https://commons.wikimedia.org/wiki/File:Window_function_(hamming).svg?uselang=es#Licencia, [Accessed 07-11-2024], 2005.

[40] A. T. Ali, H. S. Abdullah y M. N. Fadhil, "Voice recognition system using machine learning techniques," *Materials Today: Proceedings*, págs. 1-7, 2021.

[41] J. Gnanamanickam, Y. Natarajan y S. P. KR, "A hybrid speech enhancement algorithm for voice assistance application," *Sensors*, vol. 21, n.º 21, pág. 7025, 2021.

[42] A. Pradhan, A. Lazar y L. Findlater, "Use of intelligent voice assistants by older adults with low technology use," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 27, n.º 4, págs. 1-27, 2020.

[43] D. Giese y G. Noubir, "Amazon echo dot or the reverberating secrets of IoT devices," en *Proceedings of the 14th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, 2021, págs. 13-24.

[44] E. Archanco, *Diez años de Siri*, https://www.applesfera.com/general/siri-decimo-aniversario, [Accessed 07-11-2024], 2021.

[45] G. Alexakis, S. Panagiotakis, A. Fragkakis, E. Markakis y K. Vassilakis, "Control of smart home operations using natural language processing, voice recognition and IoT technologies in a multi-tier architecture," *Designs*, vol. 3, n.º 3, pág. 32, 2019.

[46] K. Hounsell, *What is a Voice Transcriber When You May Need One*, nov. de 2024. dirección: https://go.verbit.ai/blog/what-is-a-voice-transcriber-when-you-may-need-one/.

[47] J. Sutherland, *¿Cómo funciona la transcripción automática? - Sonix — sonix.ai*, https://sonix.ai/resources/es/como-funciona-la-transcripcion-automatica/, [Accessed 01-08-2024].

[48] *What Is NLP (Natural Language Processing)? | IBM — ibm.com*, https://www.ibm.com/topics/natural-language-processing, [Accessed 02-08-2024].

[49] A. Graves, S. Fernández, F. Gomez y J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," en *Proceedings of the 23rd international conference on Machine learning*, 2006, págs. 369-376.

[50] J. Tornero, "Machine Learning: Modelos Ocultos de Markov (HMM) y Redes Neuronales Artificiales (ANN)," Tesis doct., Universidad de Barcelona, jun. de 2017.

[51] *Modelo Oculto de Markov &x2013; Numerentur.org — numerentur.org*, https://numerentur.org/markov-hmm/, [Accessed 03-08-2024].

[52] J. Wall, J. Dains, J. Flynn, B. Solomon y R. Stewart, "Anamnesis y entrevista," en Octava. Baltimore, Maryland: ELSEVIER. dirección: https://clea.edu.mx/biblioteca/files/original/2e64959932bfb875e19f9a7ce139a359.pdf.

[53] D. J. Rojas, *Examen físico: inspección, palpación, percusión, auscultación*, https://escuelitamedica.com/2018/09/11/examen-fisico-inspeccion-palpacion-percusion-auscultacion/, [Accessed 07-11-2024], 2018.

[54] S. K. Pharaon, S. Schoch, L. Marchand, A. Mirza y J. Mayberry, "Orthopaedic traumatology: fundamental principles and current controversies for the acute care surgeon," *Trauma surgery & acute care open*, vol. 3, n.º 1, e000117, 2018.

[55] M. Shaw, A. M. Pelecanos y A. M. Mudge, "Evaluation of internal medicine physician or multidisciplinary team comanagement of surgical patients and clinical outcomes: a systematic review and meta-analysis," *JAMA Network Open*, vol. 3, n.º 5, e204088-e204088, 2020.

[56] A. Aggarwal, "The evolving relationship between surgery and medicine," *AMA Journal of Ethics*, vol. 12, n.º 2, págs. 119-123, 2010.

[57] M. A. Matos, L. G. Lima y L. A. A. de Oliveira, "Predisposing factors for early infection in patients with open fractures and proposal for a risk score," *Journal of Orthopaedics and Traumatology*, vol. 16, págs. 195-201, 2015.

[58] V. Lall e Y. Liu, "Contextual Biasing to Improve Domain-specific Custom Vocabulary Audio Transcription without Explicit Fine-Tuning of Whisper Model," *arXiv preprint arXiv:2410.18363*, 2024.

[59] M. I. Marrufo-Pérez, D. del Pilar Sturla-Carreto, A. Eustaquio-Martín y E. A. Lopez-Poveda, "Adaptation to noise in human speech recognition depends on noise-level statistics and fast dynamic-range compression," *Journal of Neuroscience*, vol. 40, n.º 34, págs. 6613-6623, 2020.

[60] Nuance, "Improved far field speech recognition with microphone arrays," *Healthcare RD*, 2022.

[61] W. Lück y P. Linnell, "Localization, Whitehead groups and the Atiyah conjecture," *Annals of K-Theory*, vol. 3, n.º 1, págs. 33-53, 2017.

[62] T. Valenta y L. Šmídl, "On the impact of sentence length on recognition accuracy," en *2014 12th International Conference on Signal Processing (ICSP)*, 2014, págs. 500-504. DOI: 10.1109/ICOSP.2014.7015055.

[63] A. Bodepudi, M. Reddy, S. Gutlapalli y M. Mandapuram, "Voice Recognition Systems in the Cloud Networks: Has It Reached Its Full Potential?" *Asian Journal of Applied Science and Engineering*, vol. 8, págs. 51-60, mayo de 2019. DOI: 10.18034/ajase.v8i1.12.

[64] K. Kuhn, V. Kersken, B. Reuter, N. Egger y G. Zimmermann, "Measuring the accuracy of automatic speech recognition solutions," *ACM Transactions on Accessible Computing*, vol. 16, n.º 4, págs. 1-23, 2024.

[65] B. Worthy, "Word error rate mechanism, ASR transcription and challenges in accuracy measurement," *GMR Transcription,[Online]. Available: https://www. gmrtranscription. com/blog/word-error-ratemechanism-asr-transcription-and-challenges-in-accuracymeasurement.[30 April 2021]*, 2019.

[66] A. Zhang, N. Ballas y J. Pineau, "A dissection of overfitting and generalization in continuous reinforcement learning," *arXiv preprint arXiv:1806.07937*, 2018.

[67] I. of Data, "Measuring Success: Techniques for Evaluating Classification Models in Data Science," *Analytics, Data Analysis, Data Science*, 2024.

[68] N. H. Shah, A. Milstein y S. C. Bagley PhD, "Making Machine Learning Models Clinically Useful," *JAMA*, vol. 322, n.º 14, págs. 1351-1352, oct. de 2019, ISSN: 0098-7484. DOI: 10.1001/jama.2019.10306. eprint: https://jamanetwork.com/journals/jama/articlepdf/2748179/jama_shah_2019_vp_190104.pdf. dirección: https://doi.org/10.1001/jama.2019.10306.