Detección de Alzheimer mediante Análisis Estilístico y de Contenido del Habla

Por:

Carlos Antonio Olachea Hernández

Tesis sometida como requisito
para obtener el grado de:

MAESTRÍA EN CIENCIAS EN EL ÁREA DE CIENCIAS COMPUTACIONALES

en el

Instituto Nacional de Astrofísica,
Óptica y Electrónica

Agosto, 2023

Tonantzintla, Puebla

Dirigida por:

Dr. Luis Villaseñor Pineda

Dr. Manuel Montes y Gómez

Agradecimientos

Este trabajo de investigación fue llevado a cabo gracias al apoyo brindado por el Consejo Nacional de Humanidades, Ciencia y Tecnología (CONAHCYT), a través de la beca del programa de Maestría en Ciencias en el Área de Ciencias Computacionales 001640.

Asimismo, quiero expresar mi gratitud a mis asesores, Dr. Luis Villaseñor Pineda y Dr. Manuel Montes y Gómez, por guiarme a lo largo de esta etapa. Sus comentarios y consejos fueron valiosos para alcanzar las metas planteadas en este trabajo.

Agradezco a mis sinodales, Dra. Delia Irazú Hernández Farías, Dr. Gustavo Rodríguez Gómez y Dr. Saúl Zapotecas Martínez por sus comentarios que me ayudaron a enriquecer este trabajo.

Al INAOE, a todo su personal y sus profesores, por la formación que me han brindado para alcanzar mis metas.

A todos mis amigos y compañeros por su ayuda incondicional y con los que compartí maravillosos momentos.

Finalmente, a mi madre, Miriam Hernández Cárdenas, por animarme a seguir estudiando y por siempre apoyarme en todos mis proyectos, Gracias.

Abstract

Currently, due to population aging, there has been an increase in Alzheimer's cases, with the most affected countries being those with aging populations. In Mexico, the prevalence of this disease in older adults is 7. 3 %, and it is expected that the number will increase with time. Alzheimer is a difficult disease to detect, which often means that it is not detected until it is at an advanced stage; this makes pharmacological treatments ineffective. In addition, affected people require multiple care, especially in the later stages where they may have limited mobility and independency. Although memory loss is its most well-known symptom, language problems can occur from the early stages; these manifest as interruptions in speech known as disfluencies. This has motivated the search for ways to diagnose Alzheimer by analyzing audios using computational models.

In previous work, this task was attempted to be solved by extracting features from the audio. However, in recent years, it has been observed that using linguistic features from audio transcripts offers good results, outperforming models that only use acoustic features. While the optimal features for solving the task are still unknown, in recent literature Transformer-based models have dominated the task. However, these models depend on the subject, which means that without major changes in semantic speech in a person, the model struggles. In addition, relevant features such as the distribution of pause and disfluency features have been omitted.

Consequently, a method is proposed to address the problem from two different perspectives, combining semantic features based on word embeddings, denoted as "content" features, with another lexical-semantic and paralinguistic group called "style" features. The method with the combined features achieves an accuracy of 0.87, outperforming both the baseline at 0.83 and the state of the art at 0.85. The results and subsequent feature analysis confirm that approaching the task through two different modalities results in a better overall classification performance. Additionally, it was detected that using style features allows for better initial detection of Alzheimer's cases.

Resumen

En la actualidad, por efectos del envejecimiento poblacional, se ha observado un incremento en los casos de Alzheimer, siendo los principales afectados países con poblaciones envejecidas. En México, esta enfermedad tiene una prevalencia del 7.8 % en adultos mayores, y se espera que el número se incremente con el tiempo. El Alzheimer es una enfermedad difícil de detectar, dado que a menudo esta detección sucede en etapas avanzadas de la enfermedad, lo cual vuelve los tratamientos farmacológicos inefectivos. Además, los afectados requieren de múltiples cuidados; especialmente en las últimas etapas, donde pueden verse limitados en su movilidad e independencia. Si bien el síntoma más conocido es la pérdida de memoria, pueden presentarse problemas del lenguaje desde etapas tempranas de dicha enfermedad; que se manifiestan en forma de interrupciones de lenguajes conocidas como disfluencias. Esto ha motivado, buscar formas de diagnosticar la enfermedad a través del análisis de audios, mediante modelos computacionales.

En trabajos previos, la tarea se intentó resolver utilizando características extraídas del audio. Sin embargo, en años recientes, se ha observado que utilizar características lingüísticas de transcripciones de los audios, ofrece buenos resultados superando a modelos que solo utilizan características acústicas. Si bien, todavía se desconoce cuáles son las características óptimas para resolver la tarea, en trabajos recientes, modelos basados en Transformers han dominado la tarea. Sin embargo, estos modelos son dependientes del tema, esto significa que si no hay cambios importantes en la parte semántica de habla de una persona, el modelo presenta problemas; además, se han omitido características relevantes como la dispersión de características como pausas y disfluencias.

Por consiguiente, se propone un método para abordar el problema desde dos perspectivas distintas, combinando características semánticas basadas en *word embeddings*, también llamadas de "contenido", con otro grupo que es léxico-semántico y paralingüístico denominado de "estilo". El método con las características combinadas alcanza una exactitud de 0.87, lo cual supera al baseline (0.83) y al estado del arte (0.85). Los resultados

y el posterior análisis de características confirman que aproximar la tarea mediante dos modalidades distintas, se traduce en un mejor desempeño en la clasificación en general. Además, de acuerdo a los resultados obtenidos identificamos que utilizar características de estilo, permite una mejor detección inicial de casos de Alzheimer.

Índice general

Índice de figuras

Índice de Tablas

Tabla de Acrónimos

Tabla 1: Acrónimos

Acrónimo	Descripción
AD	*Alzheimer's Disease*, Enfermedad de Alzheimer
ADR	*Active Data Representation*
ANOVA	*Analysis of Variance*
API	*Application Programming Interface*
ARI	*Automated Readability Index*
ASR	*Automatic Speech Recognition*, Reconomiento Automático de Voz
AWS	*Amazon Web Services*
BART	*Bidirectional Auto-Regresive Transformers*
BERT	*Bidirectional Encoder Representations from Transformers*
BLEU	*Bilingual Evaluation Understudy*
BOW	*Bag of Words*, Bolsa de Palabras
CFD	*Coincident Failure Diversity*, Diversidad de Errores Coincidentes
CHAT	*Codes for Human Analysis of Transcripts*
CLS	Clase
CNN	*Convolutional Neural Network*, Red Neuronal Convolucional
DT	*Decision Trees*, Árboles de Decisión
F1 (Clasificación)	Media Armónica

Continúa en la siguiente página

Tabla 1: Acrónimos (Continuación)

Fn (Audio)	Frecuencia Fundamental (F0), Armónicas (F1,...,Fn)
GELU	*Gaussian Error Linear Unit*
GFI	*Gunning Fog Index*
GMU	*Gated Multimodal Unit*
GRU	Dependencia Gramatical
HC	*Healthy Control*, Grupo de Control
HMM	*Hidden Marvok Model*, Modelo Oculto de Markov
IQR	*Interquartil Range*, Rango Intercuartil
KNN	*k-Nearest Neighbors*, *k* Vecinos más cercanos
LDA	*Linear Discriminant Analysis*, Análisis Discriminante Lineal
LR	*Logistic Regression*, Regresión Logística
LSP	Pares Espectrales de Linea
LSTM	*Long Short Term Memory*
MCC	*Matthews Correlation Coefficient*, Coeficiente de Correlación de Matthews
MCI	*Mild Cognitive Impairment*, Deterioro Cognitivo Leve
MFCC	*Mel Frequency Cepstral Coefficients*, Coeficientes Cepstrales de Mel
MLM	*Masked Language Model*, Modelado de Lenguaje Enmascarado
MLP	*Multi-layer Perceptron*, Perceptrón Multicapa
MMSE	*Mini Mental State Exam*
MRCG	*Multi-resolution Cochleagram*
NB	*Naive Bayes*
NLP	*Natural Processing Language*, Procesamiento del Lenguaje Natural
NSP	*Next Sentence Prediction*, Predicción de la Frase Siguiente
NUW	Número de Palabras Únicas
POS	*Part-of-Speech*, Partes de la Oración

Continúa en la siguiente página

Tabla 1: Acrónimos (Continuación)

RF	*Random Forest*, Bosques Aleatorios
RFE	Eliminación Recurrente de Características
RNN	*Recurrent Neural Network*
SEP	*separator*, separador
SPT	Tiempo de Fonación Estandarizado
SS	Habla Espontanea
SVF	Fluidez Verbal Semántica
SVM	*Support Vector Machine*, Máquina de Vectores de Soporte
TDNN	*Time-Delay Neural Network*
TF-IDF	*Term Frequency-Inverse Document Frequency*, Frecuencia de Término-Frecuencia Inversa de Documento
TPE	*Tree Parzen Estimator*
TRP	*Transition Relevance Place*
UD	Dependencia Universal
VAD	*Voice Active Detection*, Detección de Voz Activa
WER	*Word Errror Rate*, Tasa de Error de Palabra

Capítulo 1
Introducción

El Alzheimer es una enfermedad neurodegenerativa que provoca lesiones cerebrales, resultando en la pérdida de memoria y una disminución de la capacidad cognitiva. En México, afecta al 7.8 % de la población mayor de 60 años (Juarez-Cedillo et al. 2022). Actualmente, debido a que no existe una cura para esta enfermedad, los tratamientos disponibles se enfocan en aliviar los síntomas. Estos tratamientos buscan por medio de medicamentos, estimular la producción del neurotransmisor denominado acetilcolina, además de promover ejercicios de activación mental para desacelerar el deterioro cognitivo. No obstante, a medida que progresa la enfermedad, los medicamentos pierden su eficacia y el cuerpo es incapaz de generar la acetilcolina. Por consiguiente, la detección temprana del Alzheimer cobra importancia con el fin de maximizar los resultados del tratamiento.

La metodología médica para detectar esta enfermedad consiste en llevar a cabo diversos estudios para observar la capacidad cognitiva del paciente en un lapso de tiempo. Se aplican exámenes que deben repetirse varias veces y analizar cambios en su cognición, lo cual es fundamental para determinar el enfoque terapéutico adecuado. En el caso de existir una disminución entre dos evaluaciones, esta señalaría la presencia de un proceso degenerativo en curso.

Sin embargo, la metodología enfrenta diversos obstáculos como requerir de equipamiento especializado costoso, personal debidamente capacitado y procedimientos invasivos para obtener muestras de laboratorio. De ahí surge la necesidad de buscar otras alternativas menos costosas y fáciles de aplicar.

Una alternativa es el análisis de problemas de lenguaje que aparecen entre los primeros síntomas, por lo cual se convierten en valiosos marcadores para una detección temprana de la enfermedad. La pérdida de memoria junto a el declive cognitivo, es lo que ocasiona la pérdida de vocabulario y la afectación en la estructura de las frases. En un sentido taxonómico, los problemas de lenguaje se pueden categorizar en dos grupos: *disfluencias*

y *parafasias*; Estando el primero relacionado con interrupciones ocurridas en el discurso oral, y el segundo con el uso incorrecto de las palabras. En el grupo de las disfluencias se encuentran ejemplos como vacilaciones, pausas y muletillas; mientras que en las parafasias se encuentran los cambios de fonemas (“cama” con “capa”); la confusión de términos en un mismo campo semántico (“silla” con “mesa”) y el uso de circunloquios, concepto que hace referencia a la utilización de una descripción en vez del nombre de un objeto, debido a que tal nombre ha sido olvidado por el afectado.

Computacionalmente, el problema de la detección del Alzheimer puede ser abordado como un problema de clasificación, en donde, se obtiene una descripción de los problemas de lenguaje capturándolos de forma automática mediante algoritmos que procesan grabaciones con muestras de voz de la persona a diagnosticar. Dado que no se requiere obtener ningún otro tipo de muestra, a parte de la voz, este enfoque podría ofrecer una forma no invasiva y económica de apoyar en el diagnóstico del Alzheimer.

El uso de indicadores o características del habla para llevar a cabo esta tarea de detección, se ha investigado desde la década de 1990, aunque fue a partir de 2007 cuando se observó un aumento significativo en el interés y la cantidad de publicaciones sobre el tema (Vigo et al. 2022a). Los primeros trabajos se enfocaron en analizar características de audio, y en trabajos subsecuentes incluyeron el análisis de la sintaxis y la semántica de transcripciones obtenidas de los audios, ya sea mediante un transcriptor humano o automático (Vigo et al. 2022a; Yang et al. 2022; Luz, Haider, Fuente et al. 2020). Actualmente, gracias a las mejoras en los transcriptores automáticos existentes, se ha demostrado su alcance y utilidad en esta tarea (Luz, Haider, De La Fuente et al. 2021). Lo que reduce la necesidad de intervención humana y permite la automatización completa de la tarea.

En la actualidad, la detección temprana de Alzheimer sigue planteando un desafío sin resolver, con áreas de mejora que abarcan diversos aspectos. Por nombrar algunas, aún no existe una respuesta concluyente respecto a qué conjunto de características utilizar, o a qué metodología de aprendizaje máquina recurrir. En especial, cuando se desea identificar las etapas tempranas de la enfermedad de Alzheimer. Por consiguiente, es necesario explorar nuevas características y métodos para construir representaciones que permitan una distinción más precisa entre individuos afectados por el Alzheimer y aquellos que conservan su cognición. Todo esto, con un énfasis especial en reducir costos y facilitar el acceso a toda la población.

1.1. Problemática

A pesar de que la utilidad de las técnicas de aprendizaje máquina para realizar la detección de Alzheimer ha alcanzado resultados interesantes, aún enfrenta retos importantes. La metodología general consiste en extraer características de las grabaciones para identificar patrones que permitan distinguir a los pacientes con Alzheimer de los individuos sanos y realizar el diagnóstico. Como se mencionó anteriormente, aún no se tiene una respuesta clara respecto al conjunto óptimo de características a utilizar. Aunque en la literatura médica se han destacado diferentes aspectos (Bird et al. 2000) como por ejemplo la pérdida de sustantivos antes de otros tipos de unidades gramaticales, es importante considerar la posibilidad de que aún existan otras características relevantes por identificar.

Una solución comúnmente aplicada en los trabajos sobre el tema es emplear amplios conjuntos de características, principalmente obtenidas del audio. Sin embargo, esta estrategia resulta contraproducente debido a la limitada disponibilidad de datos, lo que ocasiona un sesgo en los modelos y que no puedan ajustarse apropiadamente a nuevos conjuntos de datos (Luz, Haider, Fuente et al. 2020; Martinc et al. 2020).

Otro aspecto poco estudiado en esta tarea es la identificación de pacientes con deterioro cognitivo leve (MCI[1]). En las fases tempranas, los enfermos de Alzheimer suelen recibir un diagnóstico de deterioro cognitivo leve hasta que se pueda confirmar que tal deterioro es progresivo. Las personas con MCI presentan habitualmente pequeñas dificultades en la memoria afectando su lenguaje, aunque éstas no son tan evidentes como en pacientes con Alzheimer en etapas más avanzadas. De tal forma, que las expresiones de estos casos aún se asemejan a las de individuos con una cognición normal, dificultando su discriminación. No obstante, es justo la detección del deterioro cognitivo en esta fase temprana que se lograría el mayor impacto al aplicar un tratamiento adecuado. En la mayoría de los trabajos no se busca distinguir a los individuos que se ubican en esta categoría, y si se aborda, no se compara el desempeño del método propuesto con pacientes en etapas avanzadas del Alzheimer.

Respecto a las técnicas utilizadas para la detección de Alzheimer, generalmente se han enfocado en el análisis de las pausas, y se han combinado con representaciones semánticas de las transcripciones (i.e., *embeddings*). Sin embargo, se ha estudiado poco sobre la distribución de elementos lingüísticos o pausas a lo largo del discurso. Es decir, no se ha tenido en consideración la ubicación de estos elementos durante la construcción del dis-

[1]Del inglés *Mild Cognitive Impairment*

curso. Por ejemplo, si las disfluencias predominan al inicio del discurso de un participante, esto podría sugerir que no se trata de un fenómeno generalizado, sino más bien de algo originado por otra causa, como posiblemente el nerviosismo del participante al iniciar la prueba.

Por otro lado, como se mencionó previamente, en individuos en etapas iniciales de Alzheimer no presentan una afectación severa en el uso correcto de las palabras. Por consiguiente, abordar esta tarea exclusivamente desde un punto de vista semántico resulta insuficiente, siendo esencial también considerar aspectos léxico-sintácticos y paralingüísticos para lograr discernir las sutiles diferencias entre personas con MCI y una cognición normal. En resumen, los desafíos principales surgen debido a que la tarea está en una fase de exploración para encontrar las características adecuadas. Esto ha llevado a los investigadores a probar varios métodos, algunos más exitosos que otros. En la actualidad, los enfoques que utilizan representaciones semánticas obtenidas mediante modelos de lenguaje preentrenados, como los Transformers, son los que producen los mejores resultados. Sin embargo, estos resultados solo son alcanzables en situaciones limitadas donde los pacientes abordan la misma temática (por ejemplo, describiendo la misma imagen). Por otro lado, es esencial no pasar por alto la inclusión de características que brinden una discriminación más detallada. Esto es especialmente crucial en situaciones donde las afectaciones semánticas aún no son significativas, como en las primeras etapas de la enfermedad cuando hay un deterioro cognitivo leve. Identificar y describir estos casos es lo que genera un gran interés, ya que la detección temprana aumenta las posibilidades de que un tratamiento farmacológico tenga un impacto real.

1.2. Motivación

Los modelos basados en Transformers logran construir representaciones semánticas de transcripciones, que permiten distinguir a pacientes de Alzheimer. Estos modelos obtienen resultados competitivos, e incluso sobrepasan los obtenidos al emplear únicamente la modalidad de audio. Sin embargo, la efectividad de estos modelos está limitada a que haya una temática similar. Cualquier desviación de la temática o la formulación de frases extrañas, es lo que permite distinguir a una persona con problemas cognitivos de una sana. Pero en casos iniciales, se conserva todavía la capacidad para formular frases con cierta coherencia. Por consiguiente, es esencial capturar aspectos léxico-semánticos y paralingüísticos con el objetivo de analizar la estructura y encontrar sutiles diferencias, debido

a que en etapas iniciales, es más importante observar "cómo" el paciente dice las cosas, que él "qué" fue lo que dijo. Además de incluir características que tomen en cuenta la dispersión de fenómenos como pausas y disfluencias, debido a que su ubicación podría aportar información acerca de si es un fenómeno arraigado en el habla de una persona o un evento aislado. Por lo tanto, a fin de aprovechar las ventajas que ofrece un punto de vista semántico, conocido a partir de ahora como de "contenido", y uno estructural o de "estilo", se busca desarrollar un método que combine ambas y con el fin de mejorar los resultados en la detección Alzheimer mediante audio.

1.3. Hipótesis

El discurso de personas con Alzheimer puede ser representado desde dos perspectivas: Una de estilo, compuesta por características que describen la estructura del discurso; y de contenido, que abstraen la parte semántica del mismo. Ambas perspectivas son complementarias, lo que sugiere que su combinación puede mejorar la detección de la enfermedad de Alzheimer al permitir una observación de los síntomas más completa, reduciendo la ambigüedad entre casos con una sintomatología similar.

1.4. Objetivo General

Desarrollar un método que mejore la detección de casos de Alzheimer, mediante la utilización de características de estilo y contenido extraídas de grabaciones de audio y transcripciones automáticas.

1.4.1. Objetivos Específicos

- Definir un conjunto de características de estilo a partir de audio y texto que represente la forma de comunicarse de los pacientes.
- Adaptar un método para la extracción de características de contenido a partir de texto que represente la parte semántica del habla.
- Desarrollar un método de clasificación que utilice características de estilo y contenido para realizar el diagnóstico de Alzheimer.

1.5. Alcances y limitaciones

- El método desarrollado está pensado para audios que comparten una temática en común. Audios de temáticas variadas podrían dificultar el aprendizaje, al forzar al modelo basado en Transformers a aprender diferentes contextos no relacionados.
- Relacionado con lo anterior, el conjunto de características de estilo utilizado no es apropiado para tareas de detección basadas en conversaciones, en donde, la representación debe incluir características relacionadas con los turnos entre entrevistador y participante.

1.6. Organización de la Tesis

Este trabajo de tesis está organizado en los siguientes capítulos:

- Capítulo 2 - Marco Teórico: Se describen los conceptos necesarios para la compresión de este trabajo.
- Capítulo 3 - Estado del Arte: Se presenta la revisión de los trabajos que forman el estado del arte relacionado con las diferentes metodologías para la detección de Alzheimer.
- Capítulo 4 - Método Propuesto: Se introduce el modelo propuesto, describiendo la estrategia para la extracción de características de audio y su procesamiento para realizar la detección de Alzheimer.
- Capítulo 5 - Experimentos: Se detallan los diferentes experimentos realizados con el modelo propuesto y sus diferentes configuraciones.
- Capítulo 6 - Análisis y Discusión de Resultados: Se revisan los resultados obtenidos y se presentan los diferentes estudios de relevancia estadística. Se presentan el análisis de error para determinar las instancias que provocan fallas en el modelo y las maneras de mitigar sus efectos.
- Capítulo 7 - Conclusiones y Trabajo Futuro: Se presentan los comentarios finales, contribuciones del método propuesto y el trabajo futuro.

Capítulo 2
Marco Teórico

En esta sección se describen los conceptos necesarios para comprender la metodología. Varios de estos conceptos están relacionados con términos médicos empleados en el diagnóstico de Alzheimer. Por otra parte, otros se ubican dentro del tema de aprendizaje, máquina y procesamiento de lenguaje natural.

2.1. Cognición

El origen del término cognición puede remontarse a una palabra latina que significa el "acto de conocer" y se hace referencia a los mecanismos por los cuales los animales adquieren, procesan, almacenan y actúan sobre la información del entorno. Estos incluyen la percepción, el aprendizaje, la memoria y la toma de decisiones (Lyon et al. 2021). En el caso del ser humano, engloba todas las actividades del cerebro que permiten participar en el pensamiento y comportamiento conscientes; percibir y comprender el entorno; experimentar emociones, funciones mentales que comprenden la conciencia, juicio y razonamiento. En comparación con los numerosos procesos corporales que ocurren inconscientemente, como los latidos del corazón y la respiración, los procesos cognitivos requieren de pensamiento deliberado (Graff-Radford et al. 2020).

2.1.1. *Mini Mental State Exam* (MMSE)

El estado cognitivo de un individuo puede ser representado en mediante una escala a través de diversas pruebas. Una de ellas es el *mini-mental state exam* (MMSE), la cual fue propuesta por Folstein et al. (1975) y se utiliza ampliamente en entornos clínicos. Antes del MMSE se aplicaban pruebas como *Withers and Hinton* y WAIS (Gallo 2013), con un periodo de aplicación y puntuación de 30 minutos o más. Esto es un problema a la hora de

trabajar con pacientes de edad avanzada, particularmente aquellos con delirio o síndromes de demencia, que tienen lapsos de atención cortos. En consecuencia, se propuso MMSE que se estructura en 11 preguntas y es aplicable de 5 a 10 minutos. El calificativo *mini* es porque concentra en solo los aspectos cognitivos de funciones mentales, y evita cuestionar acerca del estado de humor, experiencias mentales anormales y formas de pensamiento. La prueba se divide en dos partes, la primera basada en respuestas orales relacionadas con orientación, memoria y atención. La segunda parte se basa en medir la capacidad de nombrar objetos; de seguir órdenes orales y escritas; escribir oraciones espontáneamente y copiar polígonos complejos similares a las figuras de Bender-Gestalt (Hutt 2007). Las puntuaciones de cada parte son 21 y 9 respectivamente. En total, MMSE maneja un rango cognitivo de 1 a 30; en la tabla 2.1 se desglosa el criterio para diagnosticar a un paciente basado en su puntuación de MMSE.

Tabla 2.1: Rangos para el diagnóstico de Alzheimer de acuerdo con Zaudig (1992)

Diagnóstico	Rango MMSE	Descripción
Alzheimer	[1, 22]	La autonomía del paciente está limitada y requiere cuidado continuo en casos más avanzados. Sus síntomas tienen un impacto moderado a grave en su vida diaria (p. ej. olvida cómo llegó a cierto lugar o no reconoce personas).
Deterioro Cognitivo Leve	[23, 27]	El paciente conserva su autonomía, pero presenta algunos síntomas que tienen un bajo impacto en su vida diaria (p. ej. a veces olvida dónde dejó las llaves).
Cognición Normal	[28, 30]	El paciente es autónomo y no presentan síntomas que impacten su vida.

2.1.2. Demencia y Enfermedad del Alzheimer

La demencia es un término amplio utilizado para describir un declive en la función cognitiva, en áreas como la memoria, el pensamiento y las habilidades de comunicación, de tal forma que interfieren con las actividades diarias. La demencia es causada por una enfermedad subyacente, p. ej., Alzheimer, Parkinson, demencia por cuerpos de Lewys y demencia vascular. Cada una de estas causas se caracterizan por síntomas y una progre-

sión distinta, pero en general, todas conducen a un declive en la función cognitiva con el tiempo.

Entre el 60 % y 70 % de casos nuevos de demencia son causados por Alzheimer (*Global action plan on the public health response to dementia 2017-2025* 2017), convirtiéndola en la principal causa de demencia en adultos mayores de 65 años (Graff-Radford et al. 2020). Se caracteriza por la acumulación de placas beta-amiloides y ovillos neurofibrilares en el cerebro, la cual va generando atrofia de manera paulatina. Como consecuencia de esto, las habilidades de una persona, incluyendo la memoria, el lenguaje y las habilidades cognitivas se ven afectadas (Graff-Radford et al. 2020). Las personas con Alzheimer pueden tener dificultades para comunicarse, reconocer objetos, comportarse y cumplir con necesidades físicas y básicas como comer o ir al baño. En las etapas finales de la enfermedad, los afectados pueden quedar postrados en la cama y necesitar cuidados constantes. La progresión de la enfermedad puede variar, pero generalmente acorta la vida útil de una persona. Las complicaciones que surgen de la incapacidad para moverse, comer o beber adecuadamente suelen ser la causa de la muerte. Otros factores que contribuyen a la muerte incluyen neumonía, infecciones y problemas circulatorios (Graff-Radford et al. 2020).

2.1.3. Problemas del Habla

Las personas que viven con demencia pueden experimentar dificultades con el lenguaje y recordar nombres de personas, lugares y objetos que antes resultaban familiares. En etapas avanzadas, el habla puede volverse sin sentido, repetitivo e incluir términos vagos como “cosa” y “eso” (Yuan et al. 2020). También pueden tener dificultades para comprender tanto el lenguaje escrito como hablado.

Los problemas del habla pueden ser categorizados en dos grupos principales: disfluencias y parafasias. Siendo las disfluencias interrupciones del flujo normal del habla (Yuan et al. 2020; López-de-Ipiña et al. 2013; Lou et al. 2020). Algunos tipos comunes de disfluencias incluyen:

- **Pausas:** interrupciones en el habla para organizar los pensamientos.
- **Repeticiones:** Repetir una palabra o frase.
- **Muletillas:** utilizar muletillas como “um”, “ah”, “este”, etc.
- **Vacilaciones:** cuando un hablante se detiene a mitad de la frase, frecuentemente para organizar sus ideas.

- **Correcciones:** cuando un hablante se corrige a sí mismo a mitad de una frase, para corregir un error o aclarar un significado.

Por el otro lado, las parafasias son un trastorno del lenguaje caracterizado por la selección incorrecta de palabras, o bien, la producción incorrecta de fonemas dentro de una misma palabra. Existen varios tipos (Loring et al. 2015):

- **Neológica:** Sustitución de una palabra por otra sin relación fonémica ni semántica (p. ej. *rama* → *lápiz*).
- **Fonémica:** Sustituir, añadir o eliminar fonemas de una palabra (p. ej. *lápice* → *lápiz*).
- **Semántica:** Sustitución de una palabra por otra semánticamente relacionada (p. ej. *pluma* → *lápiz*)
- **Circunloquio:** Referirse a un objeto mediante su descripción (p. ej. *con lo que escribes* → *lápiz*)

Tanto las disfluencias como las parafasias pueden ser causadas por una variedad de factores, como la dificultad para recordar palabras o ideas (López-de-Ipiña et al. 2013).

2.1.4. Diagnóstico

En el pasado, el diagnóstico de Alzheimer solía ocurrir cuando los síntomas afectaban profundamente la vida y autonomía de los pacientes, incluso a veces solo tras la autopsia (Graff-Radford et al. 2020). Sin embargo, desde los 90 y 2000, se reconoció que esta enfermedad inicia con un deterioro cognitivo leve, afectando principalmente la memoria sin interferir drásticamente en las actividades diarias, además se piensa que las bases biológicas se establecen 15 años antes de los síntomas, en la etapa preclínica (Graff-Radford et al. 2020).

Siendo la principal causa de demencia, los médicos buscan distinguir el Alzheimer de otras condiciones similares. Esto implica revisar el historial médico, medicamentos, examen físico, pruebas de laboratorio y de imagen cerebral. Algunos síntomas clave a observar son:

- **Pérdida de memoria**: Se empieza con problemas de memoria a corto plazo, p. ej. olvidar las llaves, hasta olvidar a las personas en etapas finales.

- **Problemas visuales-espaciales** (Agnosia): Existe una dificultad para detectar patrones y reconocer objetos u personas.

- **Problemas lingüísticos**: Aparecen problemas para recordar el uso correcto del lenguaje, es común observar pérdida de lenguaje y confusión entre términos.

- **Ejecución de tareas**: Se pierden habilidades y conocimientos que el afectado antes dominaba.

- **Cambios conductuales**: Es común observar irritabilidad y confusión. Tampoco es extraña la aparición de depresión y trastornos de ansiedad.

Es importante recalcar que todos los síntomas no aparecen de forma repentina. Pero en el caso de hacerlo, o bien si los cambios iniciales están más relacionados con el comportamiento que la memoria, se podría tratar de otra enfermedad neurodegenerativa. Se pueden realizar pruebas de estado mental y neuropsicológicas para evaluar la cognición, además de un examen neurológico para funciones físicas. Pruebas de laboratorio y de imagen (como resonancia magnética) ayudan a confirmar. Tras una evaluación integral y si no hay explicación alternativa, un diagnóstico de demencia por Alzheimer es probable. Detectarla temprano y brindar seguimiento mejora la calidad de vida del paciente y cuidadores.

Importancia del diagnóstico

La enfermedad de Alzheimer reduce la acetilcolina la cual es una sustancia química vital para el aprendizaje y la memoria, por lo tanto, con el objetivo de prevenir tal reducción se administran inhibidores de la colinesterasa. El tratamiento permite a los afectados conservar la comunicación cerebral y retrasar los problemas de memoria de forma temporal (Graff-Radford et al. 2020), sin embargo, en etapas avanzadas el cuerpo pierde la capacidad de producir la acetilcolina. De ahí la urgencia de realizar un diagnóstico temprano del Alzheimer, para maximizar la disminución de síntomas con medicamentos y permitir al paciente retener la autonomía suficiente para tener una comprensión de la enfermedad, participación en las decisiones médicas y ensayos clínicos.

2.2. Deterioro Cognitivo Leve (MCI)

Como se mencionó previamente, la enfermedad de Alzheimer y otras formas de demencia no suelen desarrollarse de forma repentina o inesperada (Graff-Radford et al. 2020). En cambio, el inicio de estas condiciones tiende a ocurrir gradualmente, con síntomas leves que se vuelven más graves con el tiempo. A esta etapa en donde existen síntomas leves, pero no una causa con certeza, se le conoce comúnmente como *deterioro cognitivo leve* (MCI, en inglés) y puede ser causada por la enfermedad de Alzheimer u otros trastornos similares (Graff-Radford et al. 2020). Se caracteriza por ser una condición en la que un individuo experimenta dificultades con las habilidades cognitivas, pero no tanto que afecte sus actividades diarias (Graff-Radford et al. 2020). Si bien algunas habilidades pueden permanecer normales, la persona puede exhibir patrones notables de pérdida de memoria. Las personas con MCI suelen ser capaces de vivir de manera independiente, administrar sus finanzas, realizar tareas domésticas y conducir como lo hacían antes (Graff-Radford et al. 2020). Es importante tener en cuenta que no todos los que los afectados desarrollarán demencia, algunos individuos permanecen en esta etapa e incluso pueden volver a tener una cognición normal. Aproximadamente uno de cada diez personas con deterioro cognitivo leve desarrolla demencia cada año.

Con el avance de la investigación sobre la progresión del deterioro cognitivo, una mejor comprensión del deterioro cognitivo leve puede proporcionar información sobre el desarrollo de la demencia. Este conocimiento puede llevar a mayores posibilidades de tratar las condiciones que conducen a la demencia (Graff-Radford et al. 2020).

2.2.1. Causas

El MCI no es una enfermedad, sino más bien una colección de síntomas o síndrome que pueden ser causados por varias condiciones subyacentes (Graff-Radford et al. 2020). Las causas del MCI se clasifican en categorías incluyendo neurodegenerativas, vasculares, psiquiátricas, por medicamentos, trastornos del sueño y trastornos metabólicos:

- **Trastornos neurodegenerativos:** son condiciones que causan atrofia en el cerebro, lo que conduce a un declive en la función cognitiva con el tiempo. La enfermedad de Alzheimer, la demencia de cuerpos de Lewy y la degeneración frontotemporal son ejemplos de trastornos neurodegenerativos.

- **Condiciones vasculares:** Estas afectan los vasos sanguíneos en el cerebro, lo que conduce a una limitación en el suministro de sangre, daño celular y muerte celular. Esto se conoce como demencia vascular.
- **Condiciones psiquiátricas:** La depresión, pueden afectar la memoria, la concentración y el estado de ánimo, lo que puede provocar MCI.
- **Medicamentos:** Efectos secundarios de medicamentos que afectan la función cerebral, como los opioides y aquellos utilizados para tratar la ansiedad.
- **Trastornos del sueño:** El insomnio y la apnea del sueño, pueden provocar problemas cognitivos al afectar la calidad y duración del sueño.
- **Disturbios metabólicos:** Estos ocurren cuando los procesos necesarios para mantener la vida del cuerpo se ven interrumpidos, lo que provoca problemas cognitivos leves. Los problemas de tiroides y la deficiencia de vitamina B-12 son ejemplos de trastornos metabólicos que pueden causar MCI.

2.3. Clasificación de Texto mediante CNN

Una *red neuronal convolucional* (en inglés, CNN), es un tipo de red neuronal especializada en trabajar con datos que tienen propiedades espaciales, p. ej. texto e imágenes, y soluciona el problema de la independencia entre neuronas que existen en las capas lineales de redes neuronales tradicionales (Wu et al. 2022; Zafar et al. 2018). Estas redes se componen principalmente de capas que filtran entradas en busca de características útiles. Este proceso es llamado convolución y las capas que lo aplican como **capas convolucionales**. Visualmente, se representan como una ventana llamada **kernel** que recorre la entrada dando m pasos y aplicando una operación matemática. Los valores asignados al kernel determina que patrón va a buscar y producir una fuerte respuesta cuando lo encuentre, produciendo como resultado un **mapa de activación** (Zafar et al. 2018). En términos matemáticos, una convolución se describe como una fusión de dos señales:

$$f[x,y] * g[x,y] = \sum_{n_1=-\infty}^{\infty} \sum_{n_2=-\infty}^{\infty} f[n_1,n_2] \cdot g[x-n_1, y-n_2] \tag{2.3.1}$$

En el contexto de las redes neuronales, los parámetros a aprender en las capas convolucionales son los valores del kernel, los cuales se irán actualizando de forma automática.

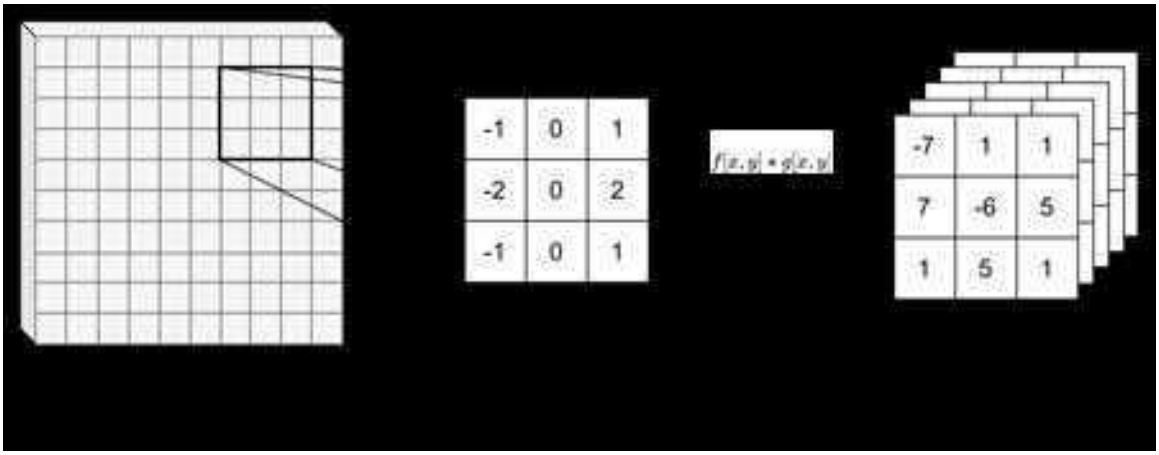

Figura 2.1: Operación de convolución mediante kernel 3x3.

También es importante rellenar las entradas con dimensiones extra para asegurar que la forma del kernel se ajuste correctamente a la forma de la entrada. Para evitar en consecuencia que el kernel genere un mapa más pequeño. En la práctica, se suele utilizar múltiples kerneles para obtener diferentes mapas de activación y capturar más patrones relevantes (Zafar et al. 2018). Para reducir los mapas de activación, se puede utilizar una **capa de agrupamiento** (o *pooling layer* en inglés). Esta capa utiliza una ventana similar al kernel, pero no tiene parámetros. En cada paso de la ventana se selecciona el valor mayor (asumiendo que es un *max pooling layer*), asignándolo a un nuevo mapa y logrando una menor dimensión. En el agrupamiento se obtiene la ventaja de que permite una reducción espacial sin afectar la profundidad (Zafar et al. 2018). Su más grande ventaja y desventaja es que al no tener parámetros no es costosa computacionalmente, pero por el otro lado, puede descartar información relevante. En general, las redes neuronales convolucionales son útiles debido a que permiten capturar patrones en el texto que de otro modo se pierden al utilizar capas lineales, debido a que estas no pueden observar ni el orden de las palabras ni si existen dependencias entre ellas. Además, que dependiendo de la representación del texto, su operación puede llegar ser costosa.

2.3.1. Extracción de Características Textuales

Al igual que las imágenes, el texto presenta propiedades espaciales, puesto que el uso y orden de las palabras puede cambiar la sintaxis y semántica de un texto. Las CNN pueden extraer relaciones entre las palabras (o tókenes), debido al uso de filtros para generar mapas de características. Los mapas representan un grupo de tókenes bajo una misma re-

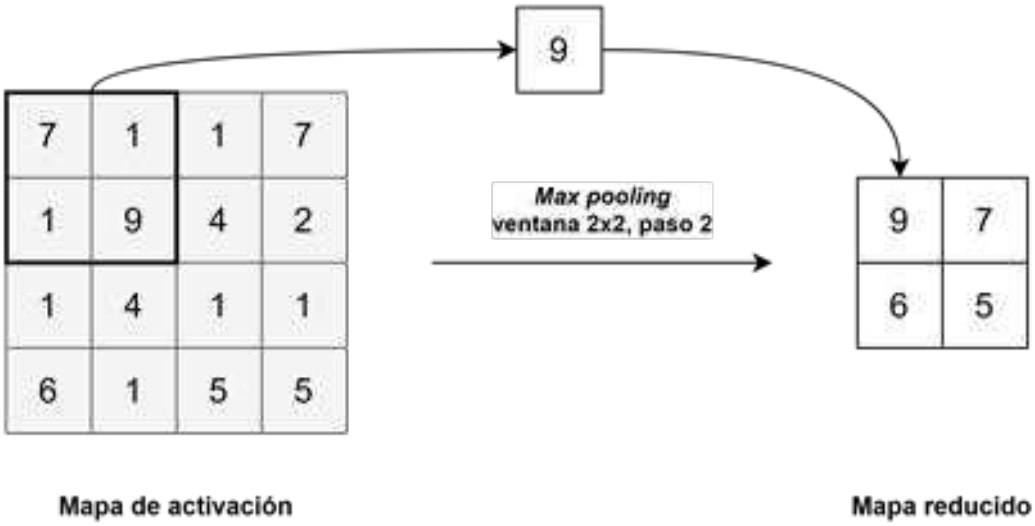

Figura 2.2: Operación de *max pooling* con una ventana de tamaño 2x2 con un paso de 2 unidades. El resultado es un mapa con los valores máximos de cada subconjunto de los datos.

presentación, es por ello que suelen considerarse un tipo de n-grama, que además pueden capturar relaciones de dependencia lingüística entre tókenes (Severyn et al. 2015).

Antes de la extracción de características, es necesario que cada token del texto esté representado de forma numérica. Esto se puede lograr de dos maneras: una es utilizar un modelo preentrenado para obtener *embeddings*; la segunda, es aprender tal representación como parte del proceso de aprendizaje. Asumiendo que se tiene una representación numérica por cada token, esto se traduce en que el texto está representado por una matriz de $n \times m$, donde n es el número de tókenes y m el tamaño de la representación. Dependiendo del problema a resolver, se tiene que fijar un filtro de tamaño $k \times m$. El valor de k, determina el número tókenes a tomar en cuenta en cada conjunto de mapas. Dicho de otro modo, si $k = 3$, entonces el filtro tomara grupos de 3 tókenes para construir la nueva representación. En la figura 2.3, se muestra la convolución de un texto dado. El filtro irá recorriendo el texto dando saltos, en un número determinado de pasos. En cada paso, se calcula el producto punto entre los embeddings del texto y el filtro, como resultado, se obtiene un escalar que representa la relación de esos 3 tókenes. La operación se repite hasta obtener un mapa de características completo. Dependiendo del problema, se puede repetir el proceso para obtener varios mapas de características para representar las diferentes relaciones en el texto. Además es recomendable utilizar filtros con diferentes

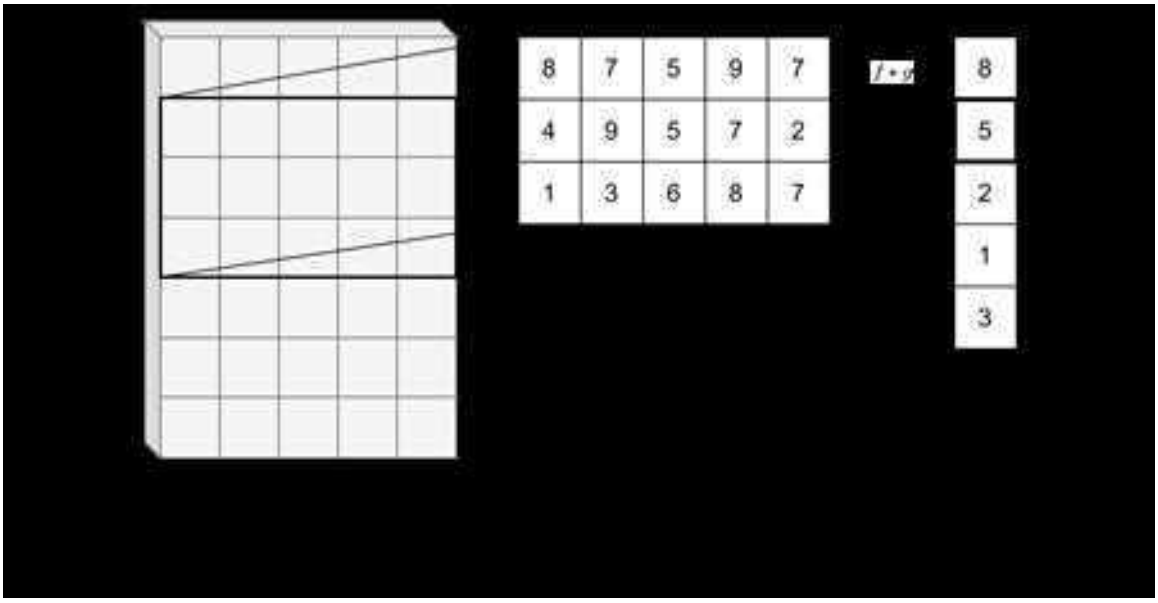

Figura 2.3: Operación de convolución con texto para obtener un mapa de características

tamaños definidos por k, de esta manera se puedan capturar relaciones de varios rangos. Una vez obtenidos los mapas de características, se pueden reducir mediante una capa de agrupamiento para obtener una sola representación de todo el texto.

2.4. Clasificación de Textos

La clasificación de textos es un proceso analítico que toma como entrada un documento y le asigna una clase a partir de un conjunto preestablecido de etiquetas (Miner et al. 2012). Este procedimiento a menudo sirve como el paso inicial en la selección de documentos para un procesamiento adicional, aunque también puede funcionar de forma independiente, como en el filtrado de spam.

La clasificación de textos implica la extracción de características para describir un documento (Miner et al. 2012). Luego, se aplica un algoritmo para analizar estas características y seleccionar la etiqueta relevante para ese documento en particular. Aunque los algoritmos de clasificación de textos suelen usar un modelo estadístico para asignar etiquetas, también se pueden aplicar enfoques basados en reglas.

Estos algoritmos evalúan varias características o rasgos de un documento, como partes de la oración, longitud, las palabras utilizadas, embeddings, etc. para determinar la categoría adecuada (Miner et al. 2012).

2.4.1. BERT

BERT (Bidirectional Encoder Representations from Transformers) es un modelo de aprendizaje profundo preentrenado que se utiliza para resolver tareas de procesamiento del lenguaje natural (NLP), como reconocimiento de entidades nombradas, respuesta a preguntas y clasificación de texto (Ravichandiran 2021). Este modelo toma como datos de entrada textos y construye una representación en forma de vectores de 768 componentes, conocidos como *embeddings*. Cada uno de estos *embeddings* abstraen información léxico-sintáctica y semántica perteneciente a un token en dentro de un contexto específico. El modelo requiere una entrada en un formato específico que consta de tres tipos de embeddings que son obtenidos del texto u oración:

- **Token embedding:** Representación numérica de un token en el texto de entrada.
- **Segment embedding:** En casos donde se proporcionan dos oraciones de diferentes segmentos como entrada, identifica el segmento correspondiente al que pertenece cada token.
- **Position embedding:** Proporciona al modelo información acerca la posición de un token dentro del texto.

Al momento de extraer los embeddings de entrada, el tokenizador de BERT inserta dos tókenes especiales:

- **CLS:** que es un token que generaliza la clase del segmento o texto, y se agrega al comienzo de cada secuencia de entrada.
- **SEP:** es un token separador utilizado para identificar el final de una oración de un segmento y el comienzo de otra oración de otro segmento.

BERT incorpora mecanismos de atención para mejorar sus resultados con datos secuenciales (Ravichandiran 2021). A diferencia de otros modelos que le precedieron, tiene la ventaja de ser bidireccional, es decir, incorpora información semántica del contexto que precede y sucede un token en específico. De hecho, una forma común de representar documentos es aplicar una operación aritmética (p. ej. suma) los embeddings de salida para agregarlos y obtener un solo embedding representa texto o documento.

La atención está implementada mediante unas unidades llamadas Transformers. La arquitectura general de un Transformer se puede ver en la figura 2.4. Dentro de estas

unidades se ubican mecanismos de atención, los cuales dado un token del texto, resaltan aquellos otros tokens con los que tiene mayor asociación o relevancia a través de un vector de pesos (Ravichandiran 2021).

La atención se calcula de la siguiente manera: Dada una matriz de entrada X de $[n, 512]$, donde n es el número de tókenes y 512 el número de componentes del embedding. Se inicializan de forma aleatoria matrices de pesos W^Q, W^K, W^V y cuyos valores finales se irán actualizando con el aprendizaje de la red. Se multiplican los pesos para obtener los vectores de *query* Q, *key* K y *value* V.

$$Q = X \circ W^Q \qquad K = X \circ W^K \qquad V = X \circ W^V \tag{2.4.1}$$

El siguiente paso es determinar el producto punto entre Q y K. La matriz resultante contiene las puntuaciones de similitud entre los tókenes de la oración, de esta forma se pueden obtener cuáles están más relacionados entre sí. Con propósito de añadir mayor estabilidad se divide entre $\sqrt{d_K}$, donde d_K es la dimensión de K:

$$\frac{Q \cdot K^T}{\sqrt{d_K}} \tag{2.4.2}$$

Se normaliza mediante Softmax para obtener una matriz de pesos:

$$\text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_K}}\right) \tag{2.4.3}$$

Para finalizar, se calcula la matriz de atención Z multiplicando la matriz de pesos con V :

$$Z = \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_K}}\right) \circ V \tag{2.4.4}$$

La matriz de atención Z corresponde únicamente a la atención de una sola cabeza; a la atención de una sola cabeza se denomina *self-attention* y a más de una *multi-head attention*.

En ocasiones, por efecto del entrenamiento, puede haber instancias en que los pesos se configuren de manera que favorezcan términos menos relevantes, lo cual ocurre en palabras que tienen un uso ambiguo (Ravichandiran 2021). Para minimizar este efecto es recomendable utilizar *multi-head attention*, que se calcula al concatenar las Z de todas las

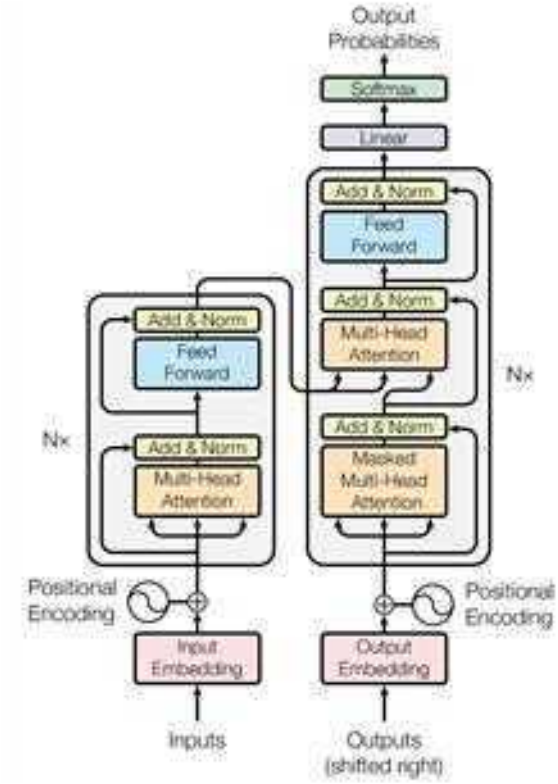

Figura 2.4: Arquitectura de Transformer

cabezas y multiplicar por una matriz de pesos W^O.

$$\text{Multi-head Attention} = \text{Concatenar}\,(Z_1, Z_2, \ldots, Z_{12})\, W^O \tag{2.4.5}$$

Para aprender todos los parámetros del modelo, se realiza un entrenamiento no supervisado mediante la resolución de dos tareas: *masked model language* (MLM) y *next-sentence prediction* (NSP) (Ravichandiran 2021). La primera es una tarea que implica el enmascaramiento de forma aleatoria de un pequeño porcentaje de tókenes de entrada y el modelo tiene que predecir la palabra oculta de acuerdo a su contexto. La tarea se realiza de forma bidireccional, es decir, que toma en cuenta los tókenes a la izquierda y derecha del token a predecir. Esto lleva a una comprensión más sólida de la entrada al tener en cuenta el contexto completo. En cuanto al enmascaramiento, se realiza normalmente de forma aleatoria con el 15 % de los tókenes de entrada. Para evitar sesgo, se introduce de forma controlada ruido durante el enmascaramiento, p. ej. en la frase “My dog is hairy”, la última palabra puede ser enmascarada de la siguiente manera:

- el 80 % de las veces se reemplaza con el token [MASK]: “My dog is [MASK]”
- el 10 % de las veces se reemplaza con otra palabra: “My dog is apple”
- el 10 % de las veces se deja sin cambios: “My dog is hairy”

La segunda tarea, *next-sentence prediction*, tiene como objetivo que el modelo que comprenda las relaciones entre oraciones (Ravichandiran 2021). NSP implica una clasificación binaria sencilla, p. ej. dadas dos oraciones A y B, esta función determina si B sigue inmediatamente a A en el documento original. La salida de la función se coloca en el token [CLS] como IsNext o NotNext, donde IsNext representa una entrada positiva y NotNext representa una entrada negativa. Para preentrenar el modelo, se utilizan una cantidad igual de muestras positivas y negativas. Al resolver MLP y NSP se aprenden patrones del contexto, y que a diferencia de los embeddings de otros modelos como word2vec (Mikolov et al. 2013), no son estáticos. Cada palabra puede tener más de una representación dependiendo de otras palabras a su alrededor. Lo que añade contenido semántico a los embeddings (Ravichandiran 2021).

Un aspecto importante es que BERT tiene la capacidad de que realizar pequeños ajustes a sus parámetros preentrenados sin la necesidad de comenzar desde cero, permitiendo que pueda adaptarse a tareas en dominios más específicos. Este proceso es conocido en inglés como *fine-tuning*. Involucra modificar los pesos de la red junto a un clasificador acoplado en la salida del modelo. Este procedimiento es menos costoso que entrenar desde cero BERT, el cual requirió 96 horas en 64 chips TPU2.

En su implementación original, el clasificador es un *multi-layer perceptron* y se entrena para clasificar el token de CLS (Ravichandiran 2021). Pero, si así se desea, una forma alternativa de utilizar BERT es omitir el clasificador, y directamente extraer la salida de la última capa de Transformers, con el fin de obtener embeddings de los tókenes. Esto permite obtener representaciones de cada token del texto, para su posterior uso en tareas de clasificación o agrupamiento.

2.4.2. Máquina de Vectores de Soporte

Las máquinas de vectores de soporte (en inglés, SVM[1]) son una familia de clasificadores que a veces son referidos como clasificador de margen óptimo, cuya estrategia general de clasificación se basa en la separación de las clases mediante hiperplanos (L.

[1]Del inglés, *support vector machine*.

Wang 2004). Asumiendo que se tiene un conjunto de l muestras etiquetadas $(\mathbf{x}, \mathbf{y})$, donde $x_i \in \mathbb{R}^n, y_i \in \{-1, 1\}$ para $1 < i < l$. Entonces, los datos de entrenamiento son separables si hay un hiperplano $H_{w,b}$ definido por un vector $w \in \mathbb{R}^n$ y un umbral $b \in \mathbb{R}$ tal que $y_i = \text{sgn}\left(w^T x_i + b\right)$ para todo $1 < i < l$ donde:

$$\text{sgn}(s) = \begin{cases} 1, & \text{Si } s \geq 0 \\ -1, & \text{Si } s < 0. \end{cases} \tag{2.4.6}$$

Si los datos son separables, entonces va a haber un número infinito de hiperplanos que los separen.

La solución al problema es encontrar aquel que maximiza el margen, es decir, la distancia mínima entre los datos y el hiperplano:

$$x_{i_p} = x_i - \gamma_i \left(w/\|w\|\right) \tag{2.4.7}$$

donde $x_{i_p} = x_i$ es la proyección normal en el hiperplano óptimo H y γ_i una medida algebraica del margen.

Dado que $x_{i_p} = x_i$ yace en H, satisface:

$$\begin{aligned} w^T x_{i_p} + b &= 0 \\ w^T \left(x_i - \gamma_i \left(w/\|w\|\right)\right) + b &= 0 \end{aligned}$$

Resolviendo γ_i y realizando una corrección para considerar casos negativos se obtiene:

$$\gamma_i = y_i \left[\left(w^T x_i + b\right)/\|w\|\right] \tag{2.4.8}$$

sujeto a:

$$\begin{aligned} w^T x_i + b &\geq 1, \text{para } y_i = +1 \\ w^T x_i + b &\leq -1, \text{para } y_i = -1 \end{aligned}$$

de las restricciones se puede deducir que:

$$\gamma = y\left[\left(w^T x + b\right)/\|w\|\right] = (\pm 1)\left[\pm 1/\|w\|\right] = 1/\|w\| \tag{2.4.9}$$

por lo que el margen está dado por $1/\|w\|$. Esto implica que maximizar el margen equivale

a minimizar la magnitud de los pesos w, o minimizar sus valores cuadráticos. Por consiguiente, el problema puede ser transformado en un problema de programación cuadrática:

$$\text{mín } \frac{1}{2}\|w\|^2 \tag{2.4.10}$$

Por lo tanto, el problema es resuelto considerando el problema dual donde:

$$\max_{\alpha} \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{2.4.11}$$

sujeto a $0 \leq \alpha_i \leq C, i = 1, \ldots, l$ y $\sum_{i=1}^{l} \alpha_i y_i = 0$, donde α corresponde a multiplicadores lagrangianos y K es una función de kernel. En cuanto a la función de decisión quedaría como:

$$f(x) = \text{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i K(x, x_i) + b\right) \tag{2.4.12}$$

Se han propuesto métodos para resolver el problema de programación cuadrática (L. Wang 2004). Algunos métodos tratan de dividir el problema de programación cuadrática en subproblemas más pequeños, donde los subproblemas menores se resuelven utilizando técnicas numéricas. Por ejemplo, el algoritmo de Optimización de Minimización Secuencial (SMO, en inglés) resuelve los subproblemas mediante métodos analíticos. El algoritmo SMO utiliza poca memoria al resolver sucesivamente un problema de optimización cuadrática que involucra dos multiplicadores de Lagrange usando métodos analíticos. El algoritmo continúa operando hasta que las condiciones Karush-Kuhn-Tucker se satisfacen para todos los multiplicadores de Lagrange. En años recientes, se han desarrollado más métodos para resolver grandes problemas de SVM. Estos métodos descomponen el problema de programación cuadrática en problemas más pequeños o encuentran soluciones aproximadas al problema de programación cuadrática reduciendo su tamaño (L. Wang 2004).

2.5. Reconocimiento Automático de Voz (ASR)

Reconocimiento Automático de Voz (ASR, en inglés) se refiere a la técnica y tecnología utilizada para convertir el lenguaje hablado en una secuencia de palabras u otras unidades lingüísticas mediante algoritmos (Li et al. 2015). El diseño y desarrollo de un sistema de

reconocimiento de voz dependen de varios componentes como la representación y preprocesamiento, clases de voz, diferentes tipos de técnicas de extracción de características, clasificadores utilizados, base de datos y rendimiento del sistema (Sen et al. 2019). La figura 2.5 muestra la estructura típica de un sistema ASR.

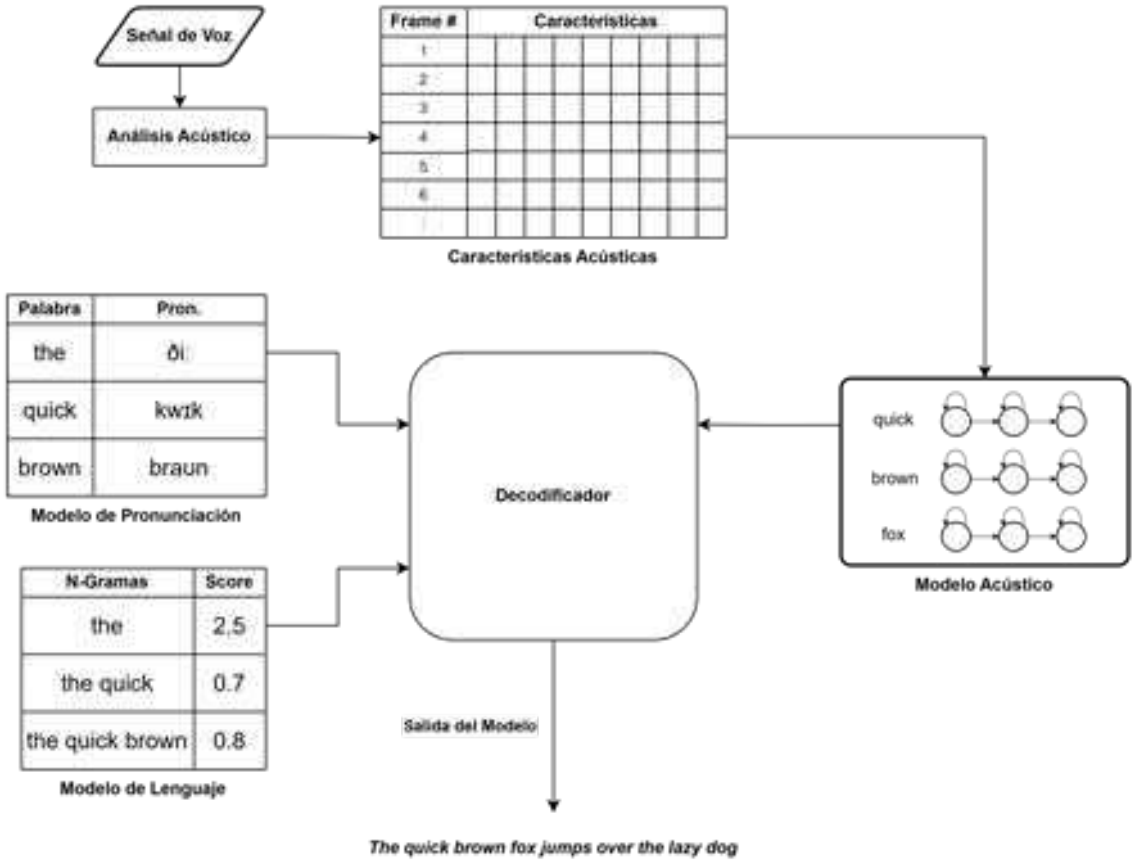

Figura 2.5: Diagrama de la estructura general de un ASR.

A partir de esta estructura general, el proceso comienza con un análisis acústico, en donde, la señal de voz es convertida a una forma discreta o en vectores de características (Sen et al. 2019). El principal objetivo es encontrar un conjunto de características o propiedades relevantes que proporcionen una representación compacta de la entrada. Para relacionar las características de audio con palabras se necesita un vocabulario. Usualmente en otros modelos de *procesamiento del lenguaje natural* (NLP), el vocabulario se compone de palabras obtenidas de un corpus, pero en sistemas de ASR si una palabra no se encuentra en el vocabulario podría generar muchos problemas durante la inferencia, por esta razón se utilizan fonemas como vocabulario. Un fonema es la parte más pequeña de

un lenguaje que forman las palabras, y que pueden distinguir una palabra de otra, p. ej. la palabra *five* puede ser separada en 3 fonemas: *f*, *ay*, *v*. Para encontrar los fonemas que componen el audio con los vectores de características se requiere de un modelo acústico, el cual puede haber sido entrenado previamente con características de una partición de datos o un conjunto diferente. Para el entrenamiento se puede aplicar algún algoritmo para crear una representación estadística de cada fonema, mediante un modelo como *Hidden Markov Model* (HMM) (Li et al. 2015). En años recientes, con el advenimiento de las redes neuronales, se han combinado con HMM obteniendo como resultado incrementos de exactitud significativos. En cuanto a la salida del modelo acústico, es un conjunto de fonemas con medidas de probabilidad. Los fonemas son mapeados a palabras mediante un modelo de pronunciación, y a su vez las palabras son ordenadas mediante el modelo de lenguaje para formar oraciones.

Lo anteriormente descrito es en términos generales, sin embargo, cada modelo tiene sus particularidades que los diferencian. Por ejemplo, en el modelo *Whisper* (Radford et al. 2023) utilizado en este proyecto prescinde del uso de un modelo de HMM, y opta por usar una arquitectura estándar de Transformer (Vaswani et al. 2017). El modelo procesa audios a 16,000 Hz, y extraen espectrogramas de Mel. Como preprocesado, esta representación es normalizada para ajustarse entre $[-1, 1]$ y es convolucionada mediante dos capas y su aplica una función de activación GELU. El resultado pasa por bloques de Transformers y en su salida final se añaden embeddings de posición sinusoidal. Además, Whisper enriquecido mediante módulos de detección de voz activa, transcriptor, alineación temporal de transcripciones, identificador de lenguaje y traductor. Lo que lo vuelve un sistema completo y robusto al ser entrenado en miles de horas de audio. En experimentos logró construir transcripciones con errores mínimos en audios con fuerte ruido ambiental.

Capítulo 3

Estado del Arte

Históricamente, las máquinas de vectores de soporte y las regresiones logísticas eran las opciones más efectivas en el uso de modelos de clasificación, dado que obtenían los mejores resultados (Vigo et al. 2022a). Debido a las limitaciones en el número de instancias, el uso de modelos profundos parecía poco práctico. Sin embargo, estudios recientes han demostrado que el uso de modelos basados en Transformers con transcripciones logra mejores resultados, incluso considerando las dificultades para realizar tareas de *fine-tuning* de manera estable (Vigo et al. 2022a; Yang et al. 2022; Yuan et al. 2020; Balagopalan, Eyre et al. 2020). Por lo tanto, la tendencia general actual es aprovechar el conocimiento preexistente en este tipo de modelos para obtener representaciones que permitan diferenciar de manera más precisa a personas cognitivamente sanas, de aquellas con la enfermedad de Alzheimer.

A continuación, se presentan trabajos que conforman el estado del arte, los cuales fueron agrupados en secciones con base en su principal aporte. En la sección de **Conjuntos de Datos**, como su nombre indica, se revisan los conjuntos de datos que son utilizados frecuentemente en la detección automática de Alzheimer; así como sus características. Uno de los aspectos importantes en la tarea es el uso de transcripciones de texto, las cuales puede ser obtenidas de forma manual con un transcriptor humano o un modelo de reconocimiento automático de voz. Cada una de estas formas de obtener transcripciones presentan ventajas y desventajas, por lo tanto, en la Sección **Transcripciones manuales y automáticas**, se revisan trabajos que comparan el uso de transcripciones manuales con automáticas, y se propone como mitigar las desventajas que introducen las transcripciones automáticas, dado que uno de los objetivos globales de la tarea es que se pueda realizar sin intervención humana. En la Sección **Exploración de Características**, se revisan trabajos que proponen métodos para resolver la tarea de clasificación y se propone el uso de ciertas características. Un problema importante dentro de la tarea es la falta de datos, en

consecuencia, se han buscado formas de construir de forma sintética datos nuevos para mejorar los métodos de clasificación. En la Sección Aumento de Datos, se hace la revisión de estos trabajos. Los audios típicamente empleados en esta tarea provienen de individuos a quienes se les aplica a un test donde deben describir una imagen. En la Sección Detección basada en Conversación, se menciona un método alternativo de clasificación que, a diferencia de describir imágenes, se centra en el análisis de audios de conversación entre un paciente y un entrevistador. A partir de esta interacción, se lleva a cabo la predicción.

3.1. Conjuntos de Datos

Los audios son grabaciones de participantes que realizan diversas tareas con el propósito de obtener una muestra de su voz para analizar y detectar la presencia de disfluencias y parafasias. Se utilizan tres metodologías principales para obtener los audios de los participantes: tareas de fluidez verbal y semántica (SVF), tareas de habla espontánea (SS) y lectura.

Las tareas de SVF se enfocan en evaluar las funciones ejecutivas, memoria y lenguaje, por ejemplo, solicitando al paciente que mencione una lista de nombres de animales o vegetales en un minuto sin repetirlos. Si el participante comete un error, la prueba finaliza y la puntuación se basa en la longitud de la lista de entidades mencionadas.

Las tareas de SS buscan que el participante hable y se analiza la fluidez de su discurso. Incluyen tareas como entrevistas, descripciones de experiencias o sueños, recuerdos de historias, descripciones de imágenes, o la lectura de textos literarios.

Figura 3.1: *Boston Cookie Theft*, esta imagen es utilizada en el diagnóstico de problemas cognitivos y afasia.

Con el objetivo de contar con datos para análisis, así como para crear de modelos computacionales, se han recopilado conjuntos de datos, tanto con tareas SFV como SS, en diferentes idiomas (incluyendo inglés, italiano, húngaro y español) (Yang et al. 2022; Vigo et al. 2022a). La tabla 3.1 presenta un resumen de los conjuntos de datos en inglés, que es el idioma más común para esta tarea. Es importante destacar que, en comparación con otras tareas de clasificación en el área de procesamiento del lenguaje natural (NLP), como la clasificación de tweets con miles de ejemplos, los conjuntos de datos más relevantes en el estado del arte son de pocas instancias y no superan las 1000 (Yang et al. 2022; Vigo et al. 2022a).

Tabla 3.1: Principales conjuntos de datos en inglés. Siglas: **BCT**=*Boston Cookie Theft*, **AD**=Alzheimer, **DD**=Demencia, **HC**=Grupo de Control, **MCI**=Deterioro Cognitivo Leve, **NC**=Desconocido.

Tarea	Nombre	Distribución	Transcripción
Conversación	The Carolina Corpus Conversation	AD(30); HC(30)	Si
	IVA Dataset	AD(17); HC(16)	No
	Framingham Heart Study Dataset	DD(223); MCI(309); HC(291)	No
BCT	Pitt Corpus	DD(208); HC(104); NC(85)	Si
	ADReSS	AD(78); HC(78)	Si
	ADReSSo	AD(122); HC(115)	No
	Wisconsin Longitudinal Study	AD(115); HC(839)	Parcial

El conjunto de datos en inglés más completo con transcripciones disponibles es el Pitt Corpus, que se basa en la tarea de describir una imagen comúnmente usada para el diagnóstico de afasia (figura 3.1). De este corpus se han obtenido dos subconjuntos usados en los foros de evaluación ADReSS y ADReSSo. A excepción de ADReSSo, los conjuntos de datos incluyen transcripciones de las grabaciones en formato CHAT. Este formato se basa en un protocolo (Pye et al. 1994) que establece una metodología no solo para marcar lo expresado por los interlocutores, sino también para incluir metadatos como comentarios, errores semánticos, fenómenos paralingüísticos, entre otros aspectos.

La mayoría de los trabajos presentados a continuación utilizan estos subconjuntos (o instancias seleccionadas directamente del Pitt Corpus por los investigadores) para facilitar la comparación de resultados.

3.2. Transcripciones manuales y automáticas

Diversos trabajos han demostrado que las características lingüísticas extraídas de transcripciones, ya sea de forma manual o automática, mejoran significativamente la detección de Alzheimer y demencia. La utilización de transcripciones manuales tiene la ventaja de que un transcriptor humano puede identificar características lingüísticas que un modelo de reconocimiento automático de voz no puede, y por lo general, las transcripciones manuales contienen menos errores. No obstante, la transcripción manual requiere de intervención humana, limitando su aplicación en la vida real. Por otro lado, la utilización de un modelo ASR tiene la ventaja de que no requiere de intervención humana, además pueden extraer características acústicas y realizar aumento de datos; al generar múltiples interpretaciones de los audios[1]. Sin embargo, la calidad de las transcripciones está directamente ligada con la calidad de los audios y del modelo para separar la voz del ruido. Por ejemplo, en el conjunto de datos ADReSSo las grabaciones presentan diversos errores metodológicos o descuidos al momento de registrar a los pacientes, tales como:

- Una distancia inconsistente entre la grabadora de voz y el paciente.
- Ruido ambiental, como el sonido de un ventilador o una puerta.
- Voz de personas aparte del paciente y entrevistador.

Como resultado, los modelos ASR tienen dificultad para discernir la voz del paciente del ruido de fondo. Por lo tanto, se han investigado los efectos que producen diferentes tipos de errores cometidos por ASR con el propósito de obtener una comprensión general del impacto de dichos errores y determinar si su uso es una alternativa viable, considerando tanto el error introducido como la conveniencia que aporta.

En el estudio llevado a cabo por Tóth et al. (2015), se realizó un experimento utilizando un modelo ASR con el objetivo de determinar si, a pesar de las imprecisiones introducidas, todavía era posible obtener indicadores acústicos utilizables para posterior clasificación.

[1]Hipótesis del Modelo ASR

Sobre audios en Húngaro, los investigadores identificaron ocho indicadores acústicos que podrían emplearse para detectar casos de Alzheimer y MCI. Estos indicadores incluyen la tasa de articulación, el número de fonemas por segundo, la longitud de las oraciones, la duración de los silencios y pausas, el número de silencios y pausas, y la tasa de vacilaciones. Estos indicadores se seleccionaron a fin de obtener un modelo independiente del contenido. Para abordar el problema del ruido en el reconocimiento automático del habla, se propuso un método menos sensible mediante el uso de un ASR cuya salida se presentara en forma de fonemas, incluyendo las muletillas (*filled pauses*) como un fonema especial. Esta elección se basó en el hecho de que los modelos de ASR disponibles comercialmente están diseñados para transcribir a nivel de palabras, y es posible que no capturen adecuadamente las características acústicas no verbales. Además, el habla de los pacientes con demencia presenta desafíos adicionales, como oraciones sin gramática e inflexiones incorrectas de palabras. Aunque el reconocimiento del habla espontánea de personas mayores sin un vocabulario definido aumenta los errores, no todos los tipos de errores en el reconocimiento de fonemas afectan la extracción de los indicadores acústicos seleccionados.

Aparte de los ocho indicadores, extendieron las características. Normalizaron la frecuencia de pausas silenciosas y muletillas, dividiéndolas por el número total de fonemas en la expresión. También agregaron características adicionales relacionadas con la duración de estas pausas, incluyendo la media y la desviación estándar. Observaron que el ASR a menudo confundía las muletillas con ciertos fonemas, como confundir el sonido “schwa” con la vocal [ø] o sustituir la palabra "hmm"por el fonema [m]. Para abordar esto, agregaron características para describir la distribución de estos dos fonemas, incluyendo la duración acumulativa (dividida por la duración total de la oración), el número de ocurrencias (dividido por el número de fonemas) y la media y desviación estándar de la duración del fonema. Estas modificaciones resultaron en un total de 81 características, conocidas como el conjunto de características extendido en sus experimentos. Finalmente, se agregó la edad y sexo, aunque los autores no esperan que esta información sea inferida por un modelo, se espera que sea proporcionada por separado. Con estos datos se creo un clasificador SVM y se reportan resultados usando transcripciones manuales, automáticas y extendidas con medidas F1[2] de 86.2, 82.5 y 85.3, respectivamente. Los resultados indican que las características propuestas muestran resistencia al ruido. Esto sugiere que al agregar características relacionadas con la duración de los fonemas, se pueden diferenciar fonemas similares. Como trabajo futuro, los autores proponen fusionar estas características

[2]Media armónica entre precisión y recuerdo

con las extraídas de transcripciones. Aunque esto puede comprometer la independencia del contenido del modelo y causar errores, también podría mejorar la detección de casos difíciles en donde el contenido de audio es crucial.

En Pan, Mirheidari, Reuber et al. (2020) se propuso una alternativa para disminuir los errores introducidos por los modelos ASR. Los modelos de ASR cuando realizan su proceso de transcripción, por cada fonema se consideran varios tókenes que pueden representar el sonido. Cada uno estos tókenes candidatos tienen un valor de confianza asociado, que representa la probabilidad de que ese token sea correcto. A partir de esta premisa, los autores proponen excluir cada token al que el modelo de ASR le haya asignado un valor de confianza bajo. Se omitieron pausas breves entre palabras de alta confianza con una duración menor a 0.1 segundos. Se registró información de ritmo de las grabaciones; el tiempo de inicio y finalización; el número de palabras para los segmentos de alta confianza. Éstas características se emplean como características de *ritmo tridimensionales* para cada segmento. Un modelo ASR fue implementado mediante Kaldi para obtener tanto la confianza de los segmentos, como transcripciones automáticas. El entrenamiento se realizó combinando los conjuntos de datos DementiaBank (Becker et al. 1994) y Hallamshire (Mirheidari et al. 2019). Se utilizó la correlación de Pearson para explorar cuál era el umbral de confianza óptimo, y se seleccionó un valor de 0.5. Una vez determinados los segmentos confiables, se extrajo el conjunto de características de IS10 utilizando OpenSMILE por cada uno; este conjunto fue utilizado en un trabajo previo arrojando los mejores resultados (Warnita et al. 2018), esto les permitiría interpretar mejor sus resultados. Al finalizar, cada segmento quedaría representado mediante un vector de características, en donde cada componente sería una característica de IS10. Luego, para la clasificación se entrenó un clasificador *bi-long-short term memory* (bi-LSTM) para clasificar las características acústicas. Para clasificar las transcripciones, utilizaron un modelo descrito en Pan, Mirheidari, Reuber et al. (2019): que es un tipo de red neuronal con Gated Recurrent Units con mecanismos de atención.

Debido a que la salida del modelo son embeddings, se le conectó una capa densa con Softmax para obtener clases. Los resultados de los experimentos fueron los siguientes: el valor de F1 de las transcripciones sin filtrar fue de 0.7555, mientras que con las transcripciones filtradas se logró 0.7725. Para combinar las transcripciones filtradas con características acústicas IS10, se fusionó la salida del modelo lingüistico con las características acústicas, y como resultado se obtuvo una F1 de 0.7834. Los resultados muestran que las palabras con una confianza baja pueden afectar la calidad de las transcripciones utilizadas

en la clasificación, por lo tanto, es preferible omitir tales segmentos.

En Balagopalan, Shkaruta et al. (2020), se realizó un experimento con el objetivo identificar los errores de mayor impacto en modelos ASR y planear mejoras en trabajos futuros. Este experimento consistía en introducir errores en transcripciones manuales de forma controlada, simulando tasas de error de palabra (WER) del 20 %, 40 % y 60 % mediante la modificación aleatoria de palabras seleccionadas mediante eliminación, inserción o sustitución de palabras. Adicionalmente, se empleó un modelo ASR basado en Kaldi para obtener transcripciones automáticas. En resumen, el experimento contó con transcripciones manuales, automáticas y simuladas.

Para verificar si los errores simulados son una aproximación de una salida verdadera de un ASR, se calculó la métrica BLEU[3] entre transcripciones manuales, automáticas y simuladas. Se encontró una fuerte correlación entre ambos resultados, por lo tanto, se asume que las transcripciones simuladas son similares a las automáticas. En otro experimento, se añadió ruido gaussiano a las características lingüísticas extraídas como otra forma de simular valores resultantes de extraerlas de transcripciones automáticas con cierto grado de WER. Evaluaron tanto los errores como el ruido simulado mediante una red neuronal de dos capas ocultas. La reducción en F1 de errores de eliminación, sustitución e inserción, fue de 10 %, 2.8 % y 6.3 % respectivamente. Los resultados indicaron que eliminar palabras produce el mayor impacto.

Se comparó los resultados entre las transcripciones con error simulado y automáticas, se encontró que los resultados eran similares cuando las simuladas presentaban errores de eliminación. Una explicación del impacto de este tipo de error se debe a que afecta directamente a la complejidad sintáctica, fenómenos del discurso y la riqueza léxica. Los autores sugieren, a partir de esta observación, optimizar los modelos ASR para que penalicen errores de eliminación, aunque no profundizaron en cómo realizarlo. Este experimento sugiere que la complejidad léxico-sintáctica es importante, lo cual explicaría por qué modelos más complejos que son más robustos a la eliminación de palabras (p. ej. BERT), obtienen mejores resultados.

3.3. Exploración de Características

En el estado del arte existen 3 principales formas de trabajar con características.

[3]BLEU es una métrica que evalúa la calidad de la traducción automática comparándola con traducciones humanas, en un rango de [0, 1] (Papineni et al. 2002).

- Trabajos que usan grandes conjuntos de atributos de distintos tipos.
- Trabajos que se centran en recopilar información sobre pausas y disfluencias para luego combinarlas con otra estrategia.
- Trabajos que hacen uso de embeddings obtenidos mediante modelos profundos.

3.3.1. Fusión de Características

Este enfoque se basa en extraer múltiples características de distintos tipos, con el objetivo de encontrar aquella combinación que permita realizar la clasificación. Frecuentemente, algunas características empleadas son seleccionadas de trabajos en otras tareas comunes en NLP, mientras que otras son tomadas de estudios médicos. Para evitar tener un gran número de características se emplean técnicas de reducción de dimensionalidad o selección de características.

Varios de estos trabajos utilizan conjuntos predefinidos de características acústicas (Luz, Haider, Fuente et al. 2020). Los más comunes son:

Emobase

Este conjunto de características consta de 988 características, que incluyen coeficientes cepstrales en la frecuencia de Mel (MFCC), frecuencia fundamental (F0), envolvente de F0, pares espectrales de línea (LSP) y características de intensidad. Se aplican funciones estadísticas a estas características.

ComParE

El conjunto de características INTERSPEECH ComParE 2013 consta de 6,373 características. Incluye energía, espectrales, MFCC y descriptores de nivel bajo relacionados con la voz, como la relación armónico-ruido, características de calidad de voz, suavizado Viterbi para F0, armonía espectral y nitidez espectral psicoacústica. También se calculan funciones estadísticas.

eGeMAPS

Este conjunto de características tiene como objetivo reducir la complejidad de los conjuntos de características anteriores manteniendo su eficacia. Consta de 88 características, que incluyen semitonos de F0[4], sonoridad, flujo espectral, MFCC, jitter,

[4]En acústica, F0 es la frecuencia fundamental armónica (440 Hz); F2, . . . , Fn donde n es un múltiplo de F0. Dependiendo del contexto, F0 y F1 son equivalentes.

shimmer, frecuencias de formantes (F1, F2, F3), relación alfa, índice de Hammarberg, características de pendiente V0 y sus funciones estadísticas.

MRCG

Las características de Multi-resolution Cochleagram (MRCG) se basan en cochleagrams, que imitan los filtros auditivos humanos. Este conjunto de características incluye 768 características por frame, derivadas de cuatro niveles diferentes de resolución. Las características capturan la distribución de energía e información espectro-temporal. Se aplican funciones estadísticas a las características de MRCG, lo que da un total de 6,912 características.

El uso de estos conjuntos predefinidos se puede observar en el *baseline* de ADReSS (Luz, Haider, Fuente et al. 2020), donde se llevó a cabo un experimento que comparaba características lingüísticas y acústicas. Se utilizaron los conjuntos de características predefinidos mencionados anteriormente, lo que resultó en la extracción de un total de 88 características eGeMAPS, 988 emobase, 6,373 ComParE, 6,912 MRCG y 13 características basadas en calcular estadísticas básicas (promedio, desviación estándar, valor mínimo y máximo) de la duración y cantidad de las vocalizaciones, pausas y ritmo del habla. Se aplicó la prueba de correlación de Pearson para descartar las características acústicas que estuvieran significativamente correlacionadas con la duración (cuando $|R| > 0.2$), excepto en el conjunto de características mínimas. Como resultado, se seleccionaron 72 características eGeMAPS, 599 emobase, 3,056 ComParE y 3,253 MRCG para los experimentos de aprendizaje automático. Además, se calcularon 34 características lingüísticas como: duración, número total de enunciados, longitud promedio de las expresiones, cociente entre tipo y total de tókenes, ratio de palabras de clases abierta y cerrada, y porcentajes de 9 categorías de partes de la oración (POS).

Para la clasificación se utilizaron cinco métodos diferentes: Latent Discrimant Analysis (LDA), Decision Trees (DT), 1-Nearest Neighbor (1NN), Random Forest (RF) y Support Vector Machine (SVM). Siendo LDA el mejor, alcanzando una exactitud de 0.62 utilizando ComParE, y con las características lingüísticas 0.75.

Los resultados obtenidos en el experimento sirvieron como punto de partida para futuros trabajos y muestran dos situaciones a destacar. En primer lugar, se evidencia que las 34 características lingüísticas fueron útiles para lograr una mejor separación entre las distintas clases. En segundo lugar, se sugiere que el bajo desempeño de las características acústicas podría deberse al uso de un gran número de ellas, lo cual no necesariamente con-

duce a una mayor exactitud debido al limitado número de instancias en ADReSS (108 de entrenamiento). Por lo tanto, resulta fundamental realizar una selección de características con el fin de eliminar aquellas que introduzcan ruido o cuya varianza no aporte información útil, evitando así problemas de sobreentrenamiento al ajustar modelos a espacios de alta dimensionalidad (por ejemplo, 3096 dimensiones de ComParE).

En Martinc et al. (2020) se expandió la exploración de características útiles, considerando *embeddings* y características de legibilidad. Se utilizaron diferentes tipos de tókenes, unigramas, bigramas, tókenes de *char4grams*, tókenes de sufijos, bigramas de etiquetas POS, características de dependencia gramatical (GRA) y características de dependencia universal (UD), para generar características TF-IDF para el análisis de texto. Se utilizaron diferentes representaciones de *embeddings* para construir características UD, pero solo se encontradas relevantes aquellas hechas a partir de *doc2vec*. Además, se probaron características de legibilidad, incluyendo el *Gunning Fog Index* (GFI), *Automated Readability Index* (ARI), *SMOG Grade* y el número de palabras únicas (NUW). Para el análisis de audio, el experimento probó diferentes conjuntos de características, incluyendo el promedio de MFCC, características extraídas mediante *active data representation* (ADR) que fue probado en los conjuntos de características de emobase, ComParE y Multi-Resolution Cochleagram (MRCG), con un rendimiento deficiente que resultó en la exclusión de algunas de estas características en experimentos posteriores.

Para los experimentos de clasificación se utilizaron cuatro diferentes algoritmos: XGBoost (XGB), RF, SVM, y Logistic Regression (LR). Se exploraron diferentes configuraciones de hiperparámetros y una configuración en ensamble. El resultado fue que el modelo LR entrenado en características GFI, NUW, duración promedio de los audios, char4gram, sufijos, POS tags y UD, logró la mejor exactitud en la clasificación de 0.77 en el conjunto de prueba oficial. Por otro lado, el ensamble de modelos produjo el peor resultado con una exactitud de 0.73. Se analizó que características eran las mejores para la clasificación, y se encontraron como las mejores a las características TF-IDF, en particular UD, POS tags y char4grams; se observó también que GFI y NUW eran relevantes, lo que indica que las medidas de legibilidad aportan información útil. En cuanto a características acústicas, solo la duración fue encontrada importante, lo que tiene sentido debido a que indica la dificultad para expresarse por parte de los participantes. Los autores no encontraron los *embeddings* de doc2vec UD importantes. Un aspecto a destacar es que los modelos configurados en ensamble, entrenados con todas las características, obtuvieron resultados peores que un solo modelo que utilizó un subconjunto de características multimodal. Esto sugiere la

importancia de seleccionar únicamente las características relevantes, evitando el uso de información redundante o ruidosa que pueda comprometer el desempeño del modelo.

En un trabajo similar Rohanian et al. (2020) utilizaron embeddings de GLOVE en lugar de características obtenidas manualmente. A estos embeddings se les agregó una categoría de palabras disfluentes, determinando si la palabra es un Repair Onset Tag, Edit Term, o una palabra correcta; esto se logró de forma automática mediante un modelo. Se utilizó un conjunto de 79 características acústicas denominado COVAREP, que incluyen prosodia, calidad de voz y espectrales. Se incluyeron estadísticas de las características, como promedio, valores máximos y mínimos, desviación estándar, etc. Las características alimentaron a dos modelos Bi-LSTM que obtuvieron a su vez características profundas de audio y texto. Luego se realizó una fusión con *feed-forward highway layers with gating*, que pondera las salidas de ambos modelos para obtener una mejor representación. Se experimentó con la utilización de las características de audio, léxicas y disfluencias por separado y en diferentes combinaciones. El peor resultado lo obtuvo el modelo que solo utilizó características de audio, alcanzando una exactitud de 0.67, mientras que el modelo solo léxico alcanzó 0.71 y con las disfluencias añadidas 0.73. El mejor modelo fue el que utilizó todas las características con una exactitud de 0.79. Los resultados indican que las características de audio seleccionadas por sí mismas no son adecuadas y que difícilmente logran describir correctamente los audios. Las características léxicas representadas por GLOVE presentan problemas para lograr una correcta representación, ya que los resultados quedan por debajo del baseline. Sin embargo, el experimento sugiere que existe complementariedad entre características acústicas y textuales, y que las disfluencias ayudan a obtener mejores resultados en la tarea.

Los resultados de estos trabajos suelen ser superiores al *baseline*, presentan la desventaja de introducir ruido de manera involuntaria al considerar características poco informativas. Además, la escasez de instancias en el conjunto de datos dificulta el ajuste de un modelo con muchas características, lo cual se puede observar en otros trabajos que obtuvieron mejores resultados con menos.

3.3.2. Disfluencias

Las pausas son el aspecto más notable en las personas que padecen de Alzheimer. Cuando enfrentan dificultades para encontrar una palabra o responder, suelen recurrir a silencios prolongados, acompañados de otros fenómenos como muletillas, vacilaciones y

repeticiones de palabras. En conjunto, estos fenómenos del lenguaje se denominan disfluencias.

Frecuentemente se les cataloga como una forma de interferencia en las señales de audio, no obstante, en el contexto de la detección del Alzheimer estas señales proveen información útil. Además, es factible identificarlas mediante modelos computacionales y emplearlas como características relevantes.

Los estudios presentados a continuación adoptaron un enfoque que aprovechaban estas disfluencias para mejorar sus resultados. La principal diferencia con los estudios presentados en la sección 3.3.1 es que, en lugar de utilizar conjuntos extensos de características, se centraron únicamente en emplear disfluencias generalmente combinándolas con otro modelo.

En Zhu et al. (2021) utilizaron un modelo wav2vec (Schneider et al. 2019) para obtener transcripciones y embeddings de los audios. Se construyó un modelo intermedio en el que tomaron los embeddings acústicos y los "tradujeron" a embeddings lingüísticos de BERT. Como segunda estrategia, se utilizaron los caracteres intermedios de wav2vec para representar las pausas en las transcripciones. Esto permitió obtener información sobre la duración de las pausas, enriquecer las transcripciones con signos de puntuación, calculando su ubicación dentro del texto mediante la aplicación de ciertos umbrales antes de pasar por un modelo BERT. Los resultados fueron que la primera estrategia utilizando el traductor alcanzó una exactitud de 0.80, superior con respecto a otros métodos que emplean características acústicas. Sin embargo, el mejor resultado fue obtenido con la segunda estrategia alcanzando una puntuación de 0.83, algo que está en concordancia con (Yuan et al. 2020), en este caso la diferencia entre exactitudes se debe a los errores introducidos por wav2vec, aunque incluso vale la pena señalar que el resultado obtenido es superior al baseline de ADReSS.

En un experimento similar (Yuan et al. 2020), generó un histograma de las pausas de los archivos de audio y se codificaron en las transcripciones utilizando signos de puntuación para cada intervalo (p. ej., "."para el primer intervalo, ".."para el segundo, etc.). Se llevó a cabo el fine-tuning en BERT y en otro modelo basado en BERT denominado ERNIE (Zhang et al. 2019). Al añadir información sobre las pausas, BERT obtuvo una exactitud del 0.85, mientras que ERNIE logró el mejor resultado con un puntaje de 0.90. Aunque tanto BERT como ERNIE carecen de una representación para las pausas, al utilizar signos de puntuación, se aprovecha el conocimiento gramatical para dirigir al modelo y lograr que aprenda una representación que distinga ambas clases.

Estos trabajos lograron resultados competitivos en comparación con otros que emplearon grandes conjuntos de atributos, lo que demuestra que solo se requiere seleccionar las características adecuadas para evitar que la baja cantidad de instancias en los conjuntos de datos sea un problema.

3.3.3. Embeddings

Realizar la clasificación mediante embeddings implica la utilización de modelos de aprendizaje profundo para aprender representaciones útiles para la clasificación. Aunque los trabajos presentados en la Sección 3.3.1 utilizaron embeddings, el enfoque de estos trabajos se centró en mejorar el modelo profundo en lugar de considerarlo simplemente complementario.

En Sarawgi et al. (2020) se experimentó con una representación aprendida a través de redes neuronales para detectar el Alzheimer. Se creó un modelo que utiliza tres tipos de características distintas: disfluencias, que incluyen número de palabras, intervenciones y pausas; características acústicas, mediante ComParE; e intervenciones. Estas características se analizan tanto individualmente como en combinación para evaluar su rendimiento en la detección de la enfermedad de Alzheimer. Las características disfluentes son normalizadas en términos de la longitud de audio, mientras que Las características acústicas se estandarizan y reducen a 21 características ortogonales mediante el análisis de componentes principales (PCA). Las características de intervenciones se basan en la secuencia de hablantes en la transcripción, es decir, en identificar qué segmentos pertenecen al participante y al entrevistador. Se construyeron tres modelos: dos de *Multi-Layer Perceptron* (MLP) para audio y disfluencias respectivamente, y otro de *Long-Short Term Memory* (LSTM) para intervenciones. Para obtener el mejor resultado, los modelos se configuraron en un ensamble cuya decisión final es tomada por una regresión logística, el cual logra una exactitud del 0.88.

Los autores no presentaron resultados de cada modalidad sobre el conjunto de prueba oficial, por lo tanto, no es posible saber si la reducción de características de ComParE fue útil. Pero si se puede interpretar que a diferencia del ensamble presentado en Martinc et al. (2020), los modelos basados en redes neuronales MLP y LSTM tienen la capacidad aprender relaciones entre las características, y a pesar del número reducido de instancias, logran generar mejores representaciones.

Modelos basados en atención han logrado obtener resultados del estado del arte en

otras tareas. En Balagopalan, Eyre et al. (2020) se llevó a cabo una comparación entre el rendimiento de modelos simples: SVM, RF, *naive bayes* (NB) y un modelo BERT. Se extrajeron características léxico-sintácticas y se realizó una búsqueda de hiperparámetros en los modelos simples. Para BERT, se exploraron diferentes números de épocas (de 1 a 10) para el fine-tuning y se utilizó un algoritmo para encontrar de manera dinámica la tasa de aprendizaje adecuada. Los resultados obtenidos mostraron que la SVM con características léxico-sintácticas logró una exactitud del 0.81, siendo el mejor modelo entre los simples. Por otro lado, aunque pudiera parecer contraintuitivo debido a la baja cantidad de instancias, el modelo BERT alcanzó una exactitud del 0.83, superando por un margen a todos los modelos simples. En este experimento BERT demostró un mejor rendimiento en comparación con otros modelos, sin la necesidad de extraer de forma manual características. Por lo tanto, el conocimiento previo de BERT y el fine-tuning permiten obtener una representación adecuada para la clasificación de instancias de Alzheimer.

Estos modelos superan o igualan a Sarawgi et al. (2020) sin necesidad de ajustar múltiples modelos. Sin embargo, los autores del ensamble no proporcionaron la exactitud de los modelos evaluados en cada modalidad en el conjunto de prueba, impidiendo la comparación de los modelos MLP con características de disfluencias y los modelos basados en Transformers. Aun así, los experimentos demuestran que las redes neuronales generan representaciones precisas de cada clase.

3.4. Aumento de Datos

Como se ha mencionado previamente, los conjuntos de datos utilizados en esta tarea suelen tener menos de 1000 instancias. Para mitigar esta escasez de instancias se han explorado técnicas de aumento de datos, con el fin de mejorar el ajuste de los modelos y aumentar la robustez del ruido en las transcripciones.

Los modelos ASR como parte de su funcionamiento generan hipótesis, es decir, posibles interpretaciones de un segmento de audio. Estas son agrupadas en *lattices* y se les asocia un coeficiente de confianza. Dado que es posible generar *n* hipótesis por cada audio, es una forma de realizar un aumento de datos al generarse transcripciones similares. En Pan, Mirheidari, Harris et al. (2021) se describe como utilizaron esta técnica para obtener transcripciones adicionales en busca de obtener mejores resultados. Para esto, se entrenó un ASR basado en Kaldi y utilizar *decoding lattices* para generar hipótesis con puntajes de confianza. Se utilizaron técnicas de transferencia de aprendizaje con varios conjuntos de

datos, incluyendo los conjuntos de datos LIBRISPEECH (Panayotov et al. 2015), AMI (Carletta et al. 2006), DR INTERVIEWS (Gratch et al. 2014), IVA y HALLAM (Pan, Mirheidari, Harris et al. 2021), así como 238 descripciones de Boston Cookie Theft de SHEFMAN-CT (Pan, Mirheidari, Harris et al. 2021). Se desarrolló el modelo ASR mediante el uso de una técnica de *transfer learning*, descrita en Manohar et al. (2018). En particular, se empleó un modelo previamente entrenado de *Time-Delay Neural Network* (TDNN) con una red neuronal de tipo LSTM y se utilizó para procesar nuevos conjuntos de datos durante una época de entrenamiento adicional. Se utilizaron *4-grams* con suavizado de Turing como datos de entrada. El ASR logró un WER del 8.23 % en un conjunto de datos de validación de SHEFMAN CT. Con el modelo ASR se generaron 30 hipótesis de cada audio y se obtuvieron coeficientes de confianza. Se concatenaron las últimas 3 capas de un BERT Large con la capa de confianza del modelo ASR. La salida de esta fusión se le conectó una capa de clasificación y se realizó *fine-tuning*. La idea de realizar este proceso es de hacer más robusto el modelo frente a los errores introducidos en las transcripciones. Al probar el modelo en el conjunto de prueba se alcanzó una exactitud de 0.85, siendo uno de los mejores resultados a la fecha de publicación del trabajo en los datos de ADReSSo. De igual forma, se probó con una fusión temprana de embeddings de wav2vec, sin embargo, el resultado no superó el *baseline* con 0.75. Y se probó con solo transcripciones obtenidas de wav2vec alcanzado 0.80, lo cual implica que los mejores resultados se obtuvieron utilizando solo características lingüísticas.

En Novikova (2021), se emplearon varias técnicas de aumento de datos que consistieron en la introducción controlada de ruido en las transcripciones manuales de ADReSS, con el objetivo de evaluar su influencia en los resultados de un modelo BERT. Las técnicas utilizadas incluyeron:

Back-translation

Traducir el texto original en inglés al alemán y luego volverlo a traducir al inglés para mantener la semántica y la estructura.

Sustitución de sinónimos

Reemplazar un porcentaje de palabras (del 10 % al 90 %) por sinónimos del corpus WordNet de NLTK, preservando el significado semántico.

Sustitución de palabras basada en embeddings

Utilizar embeddings preentrenados de word2vec para encontrar palabras similares para reemplazar (del 10 % al 90 % de las transcripciones originales).

Eliminación de muletillas (*filled pauses*)

Eliminar muletillas como “um” y “uh”.

Eliminación de unidades de información

Eliminar sujetos, ubicaciones, objetos y acciones de las transcripciones originales.

Se realizó *fine-tuning* a BERT y de los resultados sobre el test, se observó que ciertos tipos de alteraciones de texto, *back-translation*, la sustitución por sinónimos y la sustitución basada en embeddings, pueden generar ruido de una forma coherente en las transcripciones. BERT demostró un comportamiento robusto en cuanto a los valores de F1 y precisión cuando se sustituye hasta el 40 % de los tokens por similares basados en embeddings word2vec o sinónimos. Se mostraron resultados contrastantes en los valores de exhaustividad[5] y especificidad, en donde, la exhaustividad se mantiene o incluso aumenta entre un 4-9 %, mientras que la especificidad disminuye hasta un 21 %. La sustitución de palabras por sinónimos puede interpretarse como un aumento a nivel de complejidad léxica, debido a que el modelo es ajustado a múltiples formas de expresar el mismo significado. En personas con Alzheimer, la complejidad del lenguaje disminuye con el tiempo, por lo tanto, este método implícito de alterar la complejidad léxica de los textos parece ayudar a BERT para reducir la cantidad de errores verdaderos positivos al detectar Alzheimer. Sin embargo, reemplazar más del 40 % de las palabras originales disminuye la coherencia y afecta negativamente el rendimiento del modelo, incluyendo la exhaustividad.

No se encontraron diferencias significativas estadísticamente en cuanto al uso de muletillas ni a la eliminación de unidades de información, según los resultados de la prueba de McNemar. Esto sugiere que BERT no es afectado por la pérdida de estas características, las que inicialmente se consideraban relevantes.

Este estudio demuestra la resistencia de los modelos basados en transformers, como BERT, frente a diversos tipos de ruido. Sin embargo, al evaluar los resultados en términos de F1, la mejora es mínima, pasando de 0.83 a 0.84, al reemplazar palabras por sinónimos o eliminar unidades de información. En cuanto al resto de las modificaciones, en el mejor de los escenarios no se observaron cambios en los resultados. Por ende, su utilidad como una técnica para obtener mejores resultados en la clasificación, no queda demostrada de forma clara.

[5]La exhaustividad es el porcentaje de instancias predichas como clase positiva que realmente son positivas, mientras que la especificidad es el caso contrario, con la clase negativa.

3.5. Detección basada en Conversación

Existe otro enfoque para clasificar el Alzheimer que se basa en el análisis de audios de entrevistas entre un paciente y un entrevistador. El objetivo es evaluar el nivel de comprensión y capacidad de respuesta del paciente ante las preguntas planteadas. Esta metodología implica un mayor esfuerzo cognitivo que la descripción de una imagen y resulta más aplicable a situaciones del mundo real. En un trabajo que recae en esta categoría (Nasreen et al. 2021), se describe un experimento donde se emplean como características varios aspectos temporales e interacciónales de las conversaciones de diálogo basadas en la teoría de pragmática de Levinson (Levinson 1983). Estos aspectos incluyen pausas cortas, pausas largas, brechas, lapsos y silencios atribuibles. Las pausas cortas y las pausas largas se refieren a los silencios dentro del discurso de un individuo, siendo las pausas cortas inferiores a 1.5 segundos y las pausas largas superiores a 1.5 segundos. Estas pausas pueden ocurrir en *transition relevance place* (TRPs, por sus siglas en inglés) o no-TRPs, que son puntos en los que un cambio de turno en la conversación pasa de un hablante a otro. Un "brecha" es un silencio que ocurre en un cambio de hablante, mientras que un "lapso" es un retraso más largo en la comunicación en un TRP, después del cual el entrevistador inicia un nuevo tema. Además de estos aspectos, se consideraron otras características que codifican características generales de la interacción. Estas características incluyen el número de superposiciones (segmentos hablados simultáneamente por ambos interlocutores), longitud del turno (número de palabras por turno), ratio de control de la palabra (ratio de tiempo en el que habla el paciente en comparación con el tiempo total de habla), ratio estandarizado de pausas (ratio de palabras totales habladas por el paciente a pausas totales), ratio de fonación (ratio de tiempo total hablado por el paciente a tiempo total hablado incluyendo pausas) y velocidad del habla (número de palabras por minuto). Además, consideraron características acústicas extraídas con OpenSMILE (Eyben et al. 2010).

Los autores emplearon eliminación recursiva de características (RFE) para reducir la dimensionalidad del conjunto de características. Este proceso iterativo elimina las características con menor contribución a la precisión de clasificación hasta alcanzar el número deseado de características. Utilizando *grid search* para determinar el subconjunto óptimo de características tanto de las características acústicas como de las interaccionales. Se encontró varias características interactivas que se correlacionan con la enfermedad de Alzheimer (AD). Los lapsos, que se refieren a las dificultades para continuar los temas, están positivamente correlacionados con AD. Esto significa que los pacientes tienen

dificultades para mantener conversaciones, lo que resulta en retrasos cuando los entrevistadores intentan introducir nuevos temas. La duración del silencio atribuible también está positivamente correlacionada con AD, lo que indica que tienen períodos más largos de silencio en respuesta a preguntas. Por otro lado, la longitud de los turnos está negativamente correlacionada con AD. Los pacientes producen menos palabras en sus turnos, pero tienen una duración de turno más larga. Como resultado, también generan más silencios dentro de sus turnos en comparación con las personas sin AD. El tiempo de fonación estandarizado (SPT) y la tasa de fonación transformada (TPR) también difirieron entre los dos grupos.

Argumentando el número limitado de muestras, los autores eligieron clasificadores tradicionales de aprendizaje automático en lugar de redes neuronales para encontrar un equilibrio entre el rendimiento de clasificación, la complejidad de tiempo de ejecución y el riesgo de sobreajuste. En el estudio se utilizan tres clasificadores: LR, SVM y RF. Se emplean dos estrategias de fusión: fusión temprana y fusión tardía. En fusión temprana, los valores normalizados de las características acústicas e interaccionales se concatenan directamente; en fusión tardía, se realizan predicciones individualmente para cada conjunto de características, y los puntajes de predicción de cada clasificador se combinan utilizando un ensamble de votación suave.

Los resultados obtenidos mostraron que la fusión temprana logra la mayor precisión y recuperación tanto para las clases de AD como para no-AD al utilizar clasificadores LR y SVM. Los valores F1 de cada clase fueron: 0.90 (AD) y 0.89 (no-AD). La combinación de características interaccionales y acústicas mejora particularmente la recuperación de la clase AD alcanzando un valor de 0.93. La exactitud obtenida de 0.90 (LR y SVM) es superior al baseline del conjunto de datos empleado de 0.75 (LR) y 0.83 (SVM)

Inicialmente, los resultados parecen favorables, no obstante, es importante destacar una limitación inherente al conjunto de datos utilizado. Este conjunto consta de solo 30 instancias, lo cual es pequeño en comparación con los empleados en otros estudios, e inclusive comparándolo con su baseline que es de 38 instancias. Por lo tanto, resulta imprescindible ampliar el tamaño del conjunto de datos para determinar si esta metodología es verdaderamente generalizable a una población más amplia de posibles pacientes.

3.6. Discusión

En conclusión, al revisar diversos trabajos, se ha observado la necesidad de tomar medidas especiales al abordar esta tarea. En donde, existe una escasez de datos disponibles y se trabaja con audios de calidad regular, por consiguiente, es importante tener precaución al escoger características, ya que la falta de instancias dificulta el ajuste de modelos, de esta forma evitando tanto el sobreentrenamiento como la falta de convergencia al trabajar con demasiadas características. Este problema puede mitigarse mediante la selección de características o la reducción de la dimensionalidad a través de PCA. Sin embargo, esta última opción conlleva el riesgo de perder información si se realiza de manera inadecuada.

En cuanto a la relevancia de características, algunas han sido identificadas, p. ej. las pausas, que han sido empleadas en varios trabajos logrando resultados favorables al combinarlas con embeddings. No obstante, en dichos trabajos no se han aprovechado otras características lingüísticas para enriquecer el enfoque. Además, es importante destacar que en los diversos trabajos que utilizaron características lingüísticas no tuvieron en cuenta la distribución de pausas y disfluencias. Esto es relevante, ya que permite determinar si dicho fenómeno es un evento aislado o recurrente. Un evento aislado puede deberse al nerviosismo y no a un deterioro cognitivo, pero si se distribuye a lo largo del audio, es posible que se trate de un problema persistente al enfrentar la dificultad de realizar una tarea asignada. Si bien algunos estudios consideraron histogramas de pausas, omitieron también observar el tamaño o la cantidad de palabras involucradas en cada disfluencia. Una gran cantidad de palabras podría indicar que el paciente tuvo que corregirse debido a un error en su intervención.

Los modelos basados en redes neuronales presentan una buena capacidad para resolver esta tarea. No obstante, es necesario tener cuidado al entrenar estos modelos con pocos datos, ya que es crucial tener en cuenta la variabilidad entre entrenamientos. Algunos trabajos intentaron combinar algún tipo de embedding con otros tipos de características, pero no se ha analizado en detalle si existen diferencias en la manera en que estos representan a los pacientes.

En cuanto a los trabajos que utilizan ASR, se puede concluir que existe una disminución en el rendimiento de los clasificadores, principalmente debido a la incapacidad de transcribir muletillas y la omisión de palabras. Sin embargo, los resultados no son catastróficos y a pesar de esta pérdida de rendimiento, los modelos siguen mostrando resultados prometedores; se espera que a medida que mejoren los sistemas ASR este problema se

resuelva. Además, teniendo en cuenta que varias investigaciones se centran en mitigar y proteger los ASR del ruido, solo es cuestión de tiempo. Las transcripciones obtenidas de forma automática, aunque no son perfectas, proporcionan suficiente información para proponer diversas técnicas de clasificación. Por lo tanto, seguir trabajando con ellas nos acerca cada vez más a una implementación en la vida real.

En las fases tempranas, el Alzheimer suele recibir un diagnóstico de deterioro cognitivo leve (MCI[6]). Las personas con MCI presentan habitualmente dificultades en la memoria y ciertos fenómenos lingüísticos, aunque estos no son tan evidentes como en pacientes con Alzheimer en etapas avanzadas; las intervenciones de estos casos aún se asemejan en cierta medida a las de individuos con una cognición normal. En la mayoría de los trabajos no se busca distinguir a los individuos que se ubican en esta categoría, y si se aborda, no se compara el desempeño del método propuesto con personas con Alzheimer. Finalmente, en la mayoría de los estudios se pasa por alto la detección de casos moderados o MCI, o incluso, si se considera, se eliminan las instancias más graves que suelen ser la mayoría en los conjuntos de datos disponibles. Como resultado, los investigadores se ven obligados a trabajar con conjuntos de datos aún más pequeños al tener que balancearlos. No se ha investigado lo suficiente acerca de las características más relevantes para el MCI y el Alzheimer por separado, ni sobre el comportamiento de los modelos en lo que se refiere a estos dos grupos.

En conclusión, los problemas principales surgen debido a que la tarea se encuentra en una fase de exploración en busca de características óptimas. Esto ha llevado a los investigadores a experimentar con varios métodos, algunos con más éxito que otros. Siendo en la actualidad, los enfoques que hacen uso de una representación semántica obtenida mediante Transformers, los que ofrecen los mejores resultados. No obstante, es importante no descuidar la incorporación de características que ofrezcan una perspectiva alternativa para distinguir entre distintas instancias. Esto es especialmente relevante en situaciones donde la semántica aún no se ve fuertemente comprometida, algo que suele ocurrir en las etapas tempranas de la enfermedad. Por último, hace falta más trabajos orientados a identificar y describir los casos intermedios, ya que son estos los que generan un mayor interés, esto debido a que los casos iniciales son los candidatos a recibir tratamiento farmacológico.

[6]Del inglés, *Mild Cognitive Impairment*

Capítulo 4

Método Propuesto

El método propuesto aproxima la tarea de detección de Alzheimer, como un problema de clasificación binaria con dos clases: Alzheimer (AD) y Grupo de Control (HC). Esta clasificación implica etiquetar audios de pacientes de Alzheimer utilizando dos modalidades con características de contenido, que incluyen información semántica acerca de lo que dijo el paciente, y características de estilo, que representa cómo se expresa el paciente mediante indicadores lingüísticos y paralingüísticos. En la figura 4.1, se muestra el diagrama del método en sus etapas de entrenamiento e inferencia. El método comienza generando transcripciones de los audios mediante un modelo de reconocimiento automático de voz (ASR) llamado Whisper (Radford et al. 2023). Además, se realiza un análisis de las pausas mediante un algoritmo de detección de voz (VAD), que segmenta el audio y clasifica segmentos de voz-silencio para medir la duración y dispersión de las pausas. Se extraen las características de contenido y estilo a partir de las transcripciones, incluyendo categorías gramaticales, profundidad de árboles sintácticos y medidas de disfluencias. En cuanto a las disfluencias, son identificadas con un modelo que clasifica cada token y marca aquellos que pertenecen a una frase disfluente. Para las características de contenido, se realiza un *fine-tuning* a BERT y se toman los embeddings de los tókenes, que luego son enviados a una red neuronal convolucional (CNN) para identificar las relaciones y dependencias de largo alcance entre ellos. Finalmente, se realiza una fusión de ambas representaciones, para esto, se realizaron experimentos con diferentes métodos como: fusión temprana, una *Gated Multimodal Unit* (GMU) y selección de clasificadores.

Con el propósito de obtener resultados preliminares, se procedió a entrenar clasificadores específicos para cada modalidad. Para este fin, se empleó un *multilayer perceptron* (MLP) y un SVM respectivamente. En el caso de la fusión, también se optó por utilizar un MLP en forma similar. A continuación, en las siguientes secciones se proporcionan más detalles de los diferentes aspectos del método propuesto.

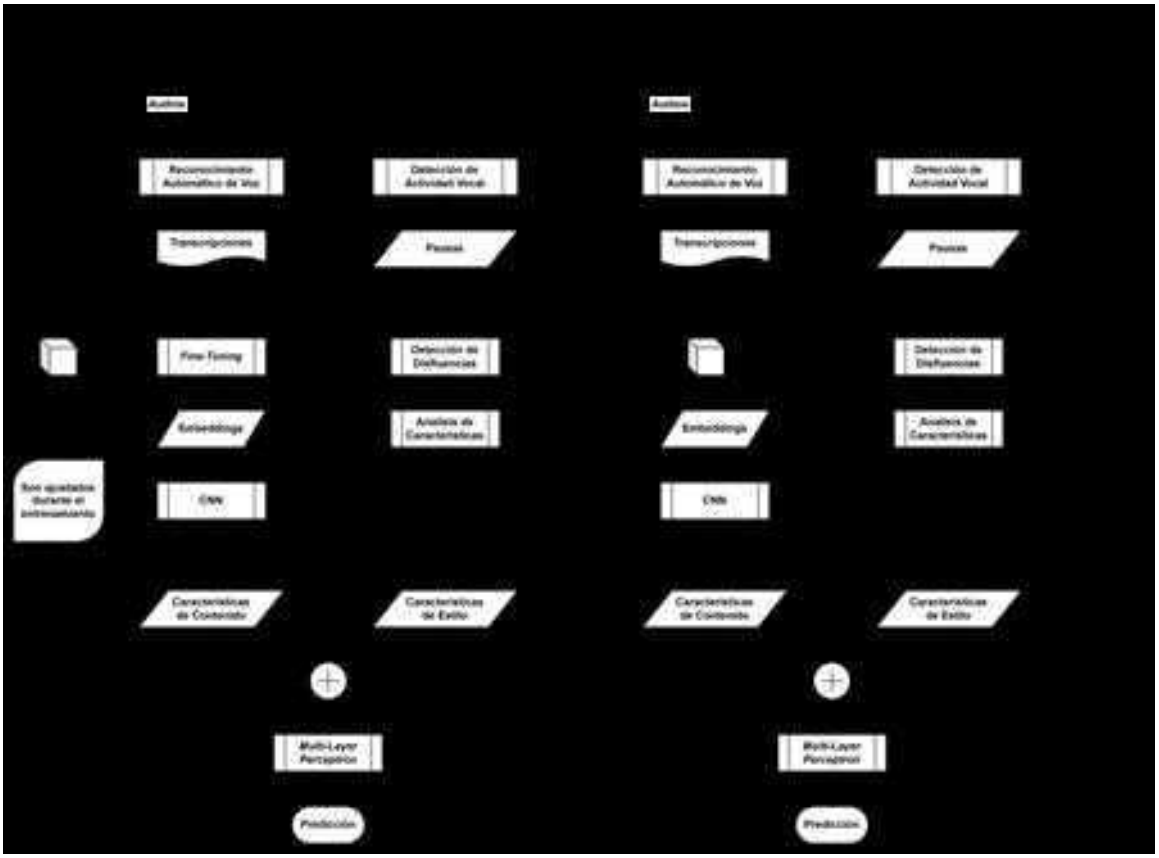

Figura 4.1: Diagrama general del método.

4.1. Características de Contenido

En esta sección, se define el conjunto de características denominadas de **contenido**, y el método con el cual son extraídas de transcripciones de audio y procesadas mediante un modelo profundo, con el fin de construir una representación de naturaleza semántica. Así cómo la configuración del clasificador utilizado en el experimento para obtener resultados de una sola modalidad.

4.1.1. Definición

Las **características de contenido** pueden ser descritas de forma coloquial como aquellas que capturan el significado de lo expresado por el participante durante su intervención, en otras palabras, el "qué" fue lo que dijo durante su participación. Especificamente, estas ofrecen una descripción semántica a partir de comprimir relaciones de dependencia

lingüística entre tókenes. Las palabras forman relaciones que en las que dependen unas de otras para formar estructuras o frases. Como resultado de la pérdida de vocabulario y otros problemas de lenguaje, es esperado que surjan patrones extraños en el uso de las palabras. Por lo tanto, se analizan las palabras enunciadas en grupos de 3, 4 y 5 tókenes, con el fin de capturar tales patrones. Un modelo de redes neuronales se encarga de tomar estos grupos y aprende una representación vectorial que resume la información de cada uno de los grupos.

4.1.2. Método de Extracción

Computacionalmente, las características de contenido se obtienen mediante la convolución de los embeddings de BERT en su versión "base". Aunque existe una versión más grande llamada BERT "large", un estudio realizado en M. S. S. Syed et al. (2020) reveló que no se encontraron diferencias significativas entre ambas versiones. Se descartan los tókenes especiales CLS y SEP, y se trabaja con los embeddings de cada token del texto. Cada embedding representa numéricamente el token en un contexto específico. Esto implica que los valores del embedding dependen de los otros tókenes que aparecen antes o después en la oración. Por lo tanto, al perderse el vocabulario, se espera encontrar patrones distintivos entre las clases, ya que se generan embeddings donde los tókenes aparecen en contextos poco comunes.

Cada embedding consta de un vector de 768 componentes. BERT tiene una restricción en la longitud de los textos que puede recibir como entrada, con un límite establecido en 300 tókenes, incluyendo los marcadores CLS y SEP. Por lo tanto, cada texto se representa como una matriz de dimensiones 298x768, que se utiliza para alimentar una CNN basada en la arquitectura propuesta en Kim (2014). Esta CNN extrae relaciones entre los embeddings y crea una nueva representación. Se emplean capas convolucionales con filtros de tamaño 3, 4 y 5 para generar diversos mapas de características que capturan dependencias de largo alcance y secuencias de contextos (Severyn et al. 2015). Cada conjunto de mapas se reduce mediante *max pooling* y se concatenan para formar un único embedding de 768 dimensiones. Dado que todos los audios describen la misma imagen, se espera que existan patrones distintivos entre los hablantes con Alzheimer y los individuos con una cognición normal.

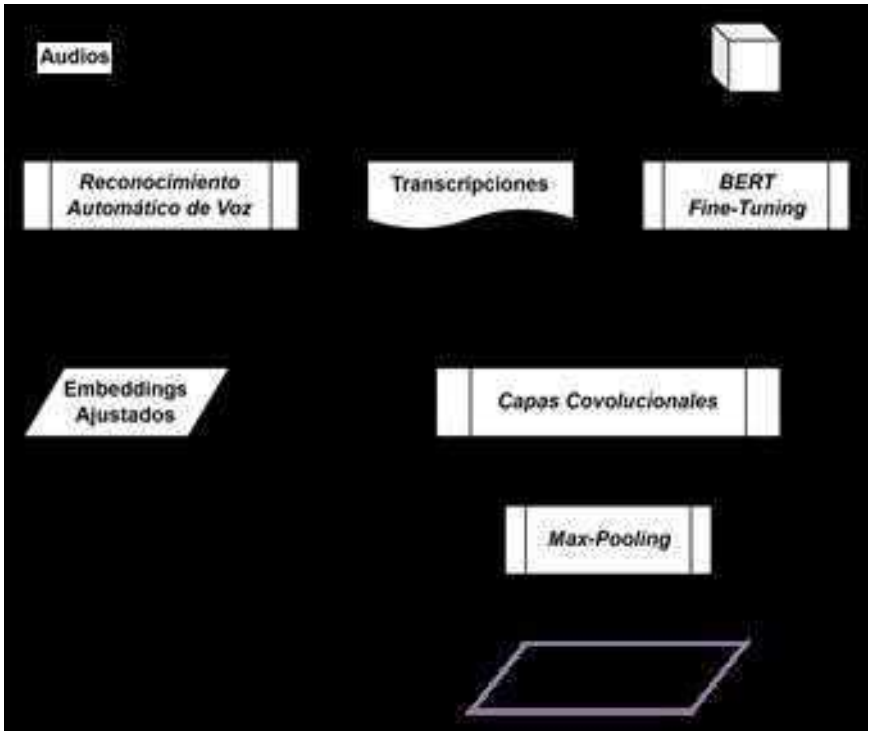

Figura 4.2: Obtención de características de contenido.

4.1.3. Clasificador

Con el fin de clasificar las instancias según sus características de contenido, se empleó un clasificador basado en una red neuronal de 3 capas. La primera capa consta de 768 unidades lineales con una activación GELU, diseñada para capturar relaciones no lineales entre los diferentes componentes de las características de contenido. La capa intermedia es una capa lineal de 768 unidades, que recibe la activación de la capa anterior y envía su salida a una capa Softmax. Esta última capa reduce las características a 2 componentes y transforma sus valores de manera que su suma total sea igual a 1. Esto se interpreta como una distribución de probabilidad entre dos clases. La predicción final se obtiene seleccionando el componente con la probabilidad más alta. La unidad 0 de la capa Softmax corresponde a la clase AD, mientras que la unidad 1 representa la clase HC. El diagrama del clasificador se muestra en la figura 4.4.

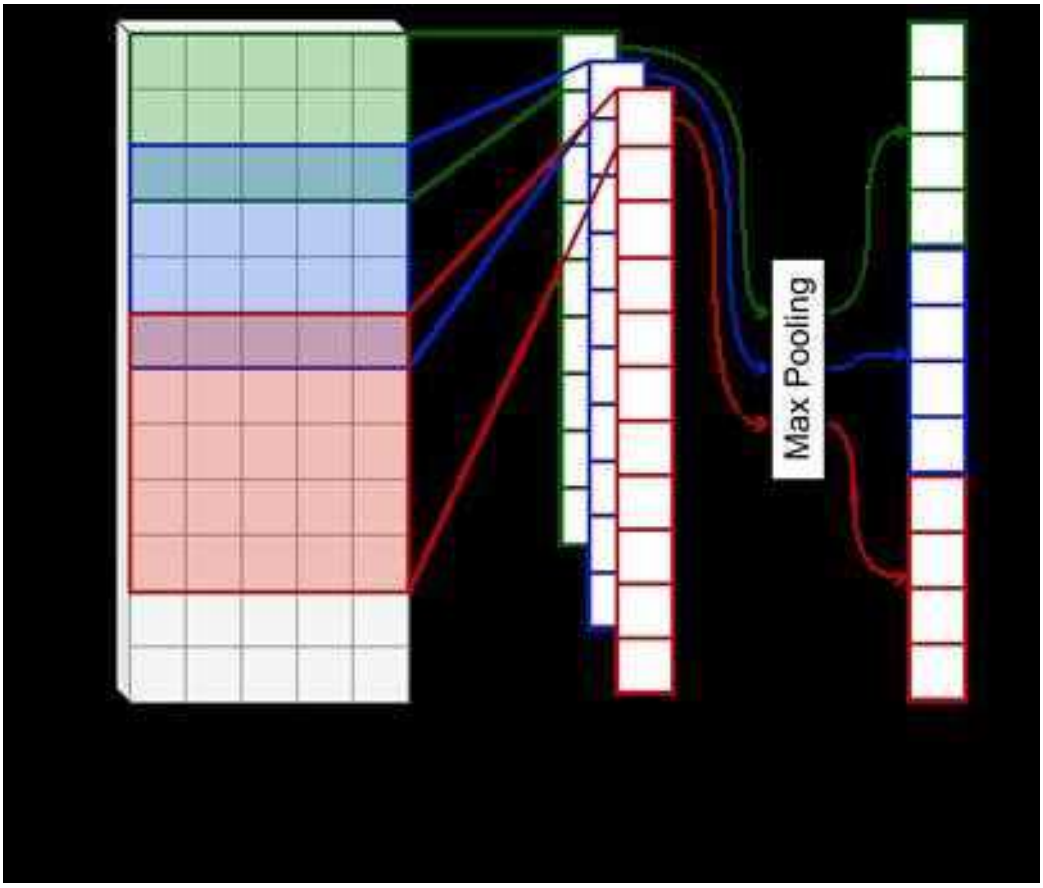

Figura 4.3: Arquitectura de la CNN.

4.2. Características de Estilo

En esta sección, se define el conjunto de características denominadas de **Estilo**, y el método con el cual son extraídas tanto del audio (en el caso de las pausas) y sus transcripciones. Para ello, se utilizaron analizadores sintácticos para etiquetar los tókenes y construir árboles de dependencia, para posteriormente extraer características relacionadas con la estructura de las frases. Además, se específica la configuración del clasificador utilizado en el experimento para obtener resultados de una sola modalidad.

4.2.1. Definición

Las **características de estilo** son un conjunto que describen la forma de comunicarse del paciente. Se diferencían de las características de contenido, en que éstas no toman en cuenta el significado de los enunciados, sino solamente su estructura. Si las características

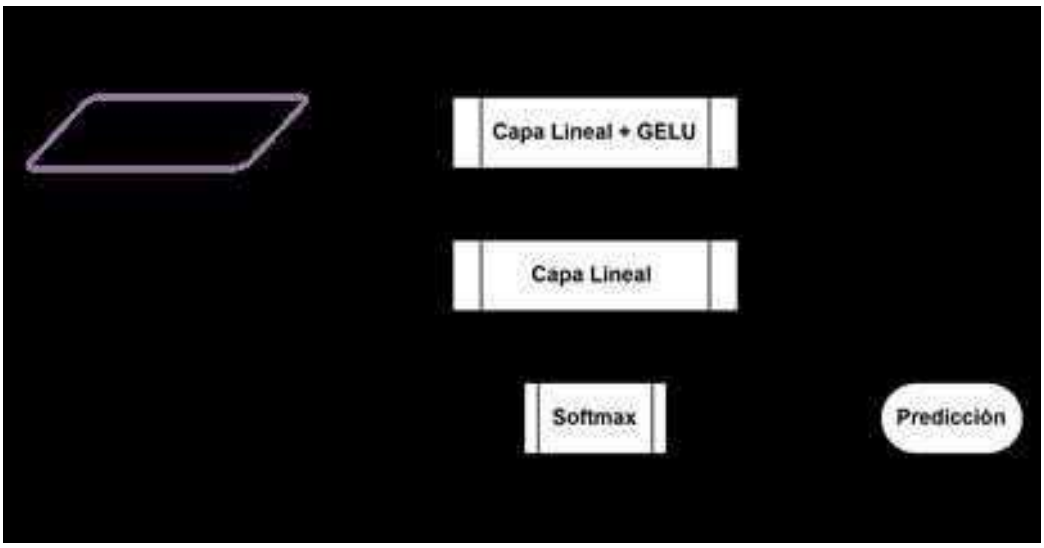

Figura 4.4: Arquitectura del Clasificador de Contenido.

de contenido corresponde al "qué", de forma análoga, las características de estilo serían el "cómo" expresó sus enunciados.

A continuación se describe cada una de las características:

Categorías Gramaticales

Con el progreso de la enfermedad de Alzheimer se ha observado que existe una pérdida no uniforme del lenguaje. Los sustantivos son los principales afectados que tienden a desaparecer primero (Bird et al. 2000). Por el otro lado, los verbos presentan mayor resistencia y se conservan inclusive las conjugaciones verbales de forma correcta. Ambas categorías tienden a desaparecer eventualmente en etapas muy tardías de la enfermedad. Por consiguiente, se utiliza un parser para el idioma inglés, con el fin de inferir la categoría de cada token y calcular su frecuencia. En total se obtienen 50 características y se enlistan en el anexo, tabla 7.1.

Profundidad media de Árbol Sintáctico

Las intervenciones de los pacientes están conformadas por múltiples árboles sintácticos. A menor profundidad indicaría que las frases tiene una estructura mucho más simple. Si en promedio las frases son más simples, mayor es la probabilidad de que se trate con alguien con pérdida de vocabulario.

Diversidad de Vocabulario

Esta métrica consiste en la cantidad única de tókenes. Se espera que una persona

cognitivamente sana tenga un vocabulario más rico que una persona con Alzheimer, que ya está sufriendo de su pérdida.

Pausas y Disfluencias

Las disfluencias son interrupciones en la fluidez del diálogo que se producen debido a la pérdida de vocabulario y disminución cognitiva. En el caso de los pacientes con Alzheimer, estas disfluencias se manifiestan en forma de pausas, muletillas, repetición de palabras y autocorrecciones.

Las pausas consisten en silencios de duración variable. Con el progreso de la enfermedad, la presencia de estas pausas tiende a aumentar, lo que las convierte en indicadores útiles para detectar rápidamente casos más graves.

Para su representación, se calcula la duración de cada pausa para construir un histograma con los intervalos: [0.5s, 1s), [1s, 2s), [2s, 5s) y [5s, ∞). Además, se incluye el número total de pausas y medidas estadísticas de sus duraciones como: duración máxima y mínima, duración promedio, desviación estándar entre la duración de las pausas y silencio total. De forma similar, para el resto de disfluencias, se calcula el número de tókenes que compone cada disfluencia. Se construye un histograma con los siguientes intervalos: [1, 3), [3, 5), [5, ∞). También se incluye el número total de disfluencias, que equivale a 14 características.

Dispersión de Pausas y Disfluencias

Es importante considerar la distribución de las pausas y disfluencias al analizar el habla de una persona. Si se concentran principalmente al comienzo del audio, es posible que indiquen un evento aislado, posiblemente debido al nerviosismo u otra razón. Por otro lado, si surgen a lo largo del audio, es probable que reflejen un fenómeno que se ha arraigado en el habla, causado por la dificultad para expresarse. Calcular la dispersión de las pausas y disfluencias es equivalente a calcular la desviación estándar muestral de las posiciones de las pausas:

$$\text{Dispersión} = \begin{cases} \sqrt{\frac{\sum (p_i - \overline{p})^2}{n-1}}, & \text{if } n > 1 \\ 0, & \text{de lo contrario} \end{cases} \tag{4.2.1}$$

donde p_i es la posición de una sola pausa; $\overline{p}$ es la posición media y n el número de pausas. En caso de que haya menos de 2 pausas o disfluencias, la dispersión es igual a 0. Lo que significa que el evento (si existe) es aislado.

Por cada intervalo de los histogramas de pausas y disfluencias anteriormente mencionados, se realiza el cálculo de la dispersión de los elementos asociados a dicho intervalo. Además, se incorpora una métrica adicional que abarca la dispersión global de todos los elementos, independientemente del intervalo al que pertenecen. Por lo tanto, se obtienen 11 características.

La representación final se obtiene al concatenar todas las características extraídas. En total, se obtienen vectores de 77 componentes por cada instancia, que en términos generales, reflejan la estructura del diálogo con la ventaja de ser independientes de la temática.

4.2.2. Método de Extracción

Las características de estilo se extraen de transcripciones, las cuales ha sido obtenidas previamente mediante un modelo automático de reconocimiento de voz. El texto es tokenizado y se emplean diversos analizadores sintácticos para capturar las características mencionadas anteriormente. En la figura 4.5 se muestra el proceso general de extracción. Por ejemplo, para determinar las categorías gramaticales, se utiliza un etiquetador de partes de la oración. En el caso de la profundidad media de los árboles sintácticos, se emplea una herramienta para construirlos, posteriormente son explorados de manera recursiva para medir su profundidad. A partir de ahí, se calcula el promedio de todas las profundidades. Finalmente, la diversidad de vocabulario equivale al número total de tókenes diferentes en los enunciados del paciente.

Para obtener información acerca de disfluencias, se utiliza un modelo del estado del arte (Lou et al. 2020) detecta secuencias de tókenes donde existen repeticiones de palabras y autocorrecciones (para más detalles, ver sección 5.2.3). Además de las características lingüísticas y disfluencias, se incluye información paralingüística sobre la presencia de pausas, detectadas en el audio mediante un modelo de detección de voz activa (para más detalles, ver sección 5.2.2). Esto permite obtener información sobre segmentos en los que probablemente hay silencio y su eventual duración. Tanto las pausas como las disfluencias se analizan en términos de su posición en los audios y transcripciones, para medir su distribución y evitar que eventos aislados y sin relevancia introduzcan errores en la clasificación.

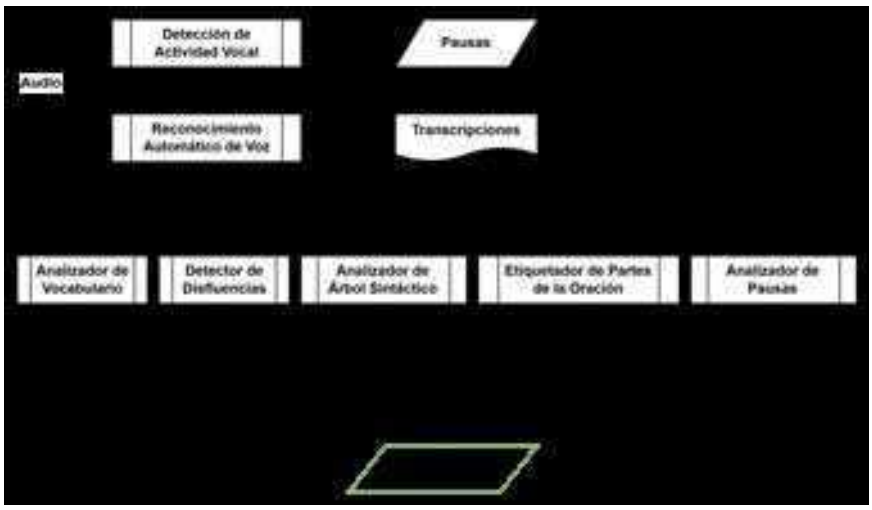

Figura 4.5: Obtención de Características de Estilo.

4.2.3. Clasificador

Para clasificar las características de estilo se utilizó un SVM, las cuales eran el estado de arte antes de la llegada de modelos basados en atención (Balagopalan, Eyre et al. 2020; Vigo et al. 2022b). Las características de estilo presentan diferentes rangos de magnitudes. Por lo tanto, los datos fueron escalados según su origen, es decir, las características que son obtenidas del análisis de las transcripciones son escaladas mediante la longitud del texto, mientras que las pausas que son obtenidas del audio, lo son por la duración en milisegundos. De forma experimental se observó que si se aplica una segunda normalización mediante IQR se obtenían mejores resultados.

La fórmula de esta normalización es la siguiente:

$$\text{Escalado IQR}(x) = \frac{x - \text{Q2}}{\text{Q3} - \text{Q1}} \tag{4.2.2}$$

donde x representa el valor de la característica a ser normalizado. $Q_3 - Q_1$ es el rango intercuartil, y Q_2 es la mediana de la población.

4.3. Fusión de Características de Contenido y Estilo

Las características de contenido ofrecen una representación semántica de lo que se dice en el audio. Por el otro lado, las características de estilo ofrecen una representación

estructural de como se dice. Al combinarse ambas representaciones, se obtiene una descripción más amplia, que no solamente permite ver si existe una coherencia en el uso de las palabras, también permite ver si hay pérdida de lenguaje, simplificación de las frases, y la presencia de fenómenos como pausas y disfluencias.

Por consiguiente, para tomar ventaja de la complementariedad entre ambos tipos de características, se exploraron tres técnicas de fusión:

Fusión temprana de características de contenido y estilo

Esta técnica implica concatenar los vectores de características de contenido y estilo, a fin de obtener una nueva representación que se envía a un clasificador. En este caso particular, es importante tener en cuenta el bajo número de instancias para lograr una fusión de características adecuada.

Fusión mediante *Gated Multimodal Unit*

Esta técnica hace uso de unas unidades denominadas *Gated Multimodal Unit* (GMU) (Arevalo et al. 2020). Originalmente, estas unidades fueron diseñadas para la fusión de información de imágenes y texto. Estas unidades reciben como entrada las características de contenido y estilo. Como hiperparámetro se requiere el número de unidades, el cuál se determinó en un número de 10 unidades de GMU. Un número mayor de unidades afecta el rendimiento o no se observan una mejora, además aumenta el tiempo de entrenamiento. La salida es una representación que combina ambos tipos de características, dando más relevancia a una "modalidad" que otra dependiendo de la entrada. Como se mencionó previamente, para algunas instancias la parte estilística tiene mayor relevancia que la parte del contenido o viceversa. Por lo tanto, en teoría, las unidades de GMU deberían ser capaces de identificar esta relevancia y representarla en su salida. Finalmente, la nueva representación se envía a un clasificador.

Fusión mediante Selección de Clasificadores

Como paso previo a la fusión, se genera un centroide para cada una de las clases: AD (Alzheimer), MCI y HC (Grupo de Control), utilizando la media de las características de estilo del conjunto de entrenamiento. Para evitar que la alta dimensionalidad afecte el cálculo de las distancias, se buscó si existía un subconjunto de características, mediante un algoritmo heurístico denominado *Tree Parzen Estimator* (TPE) (Bergstra et al. 2011).

Las mejores características encontradas fueron: sustantivos, pronombres y la dispersión de pausas en el intervalo de [1s, 2s). Cada uno de estos centroides permiten generalizar un punto de referencia, para comparar nuevas instancias y determinar qué clasificador es más confiable para ese caso en particular. Durante la etapa inferencia del conjunto de prueba, se toman las características de estilo y se compara la distancia Manhattan entre ellas y cada uno de los centroides. De esta manera, se obtienen las distancias AD_d, MCI_d y HC_d. La selección de la predicción del clasificador adecuado se realiza bajo la condición:

$$\text{Predicción} = \begin{cases} \text{C. de Estilo} & \text{Sí } AD_d < HC_d \wedge MCI_d < HC_d \\ \text{C. de Contenido} & \text{en caso contrario} \end{cases} \tag{4.3.1}$$

Es importante señalar que la cercanía a un centroide no implica que una instancia corresponda Alzheimer o MCI, solo implica que su manera de hablar (estilo) es similar. En el caso de no cumplirse con la condición, se implica que hay ambigüedad en el estilo de la persona, por lo tanto, la clasificación se realiza en términos de lo que dijo (contenido).

4.3.1. Clasificador

El clasificador que fue empleado para la fusión es un *multilayer perceptron* (MLP). Cabe destacar que este clasificador se aplica exclusivamente en el contexto de la fusión temprana y GMU. En el escenario de fusión temprana, la arquitectura del clasificador se compone de tres capas:

- La capa inicial posee 799 unidades lineales, activadas por una función GELU.
- La capa intermedia, también con 799 unidades lineales, procesa la activación de la capa anterior y su salida es enviada a una capa Softmax.
- A través de esta capa Softmax, se construye una distribución probabilística relacionada con las dos categorías: AD y HC. La categoría que registra una mayor probabilidad se designa como etiqueta para una instancia específica.

En cuanto al GMU, la estructura del clasificador se modifica, pasando de 799 unidades a solamente 10 en cada capa, con el objetivo de alinearse con la salida de la capa GMU. En

la figura 4.6 se muestra la arquitectura del clasificador. En el diagrama, el símbolo del más representa la fusión, esta puede ser mediante una fusión temprana o GMU.

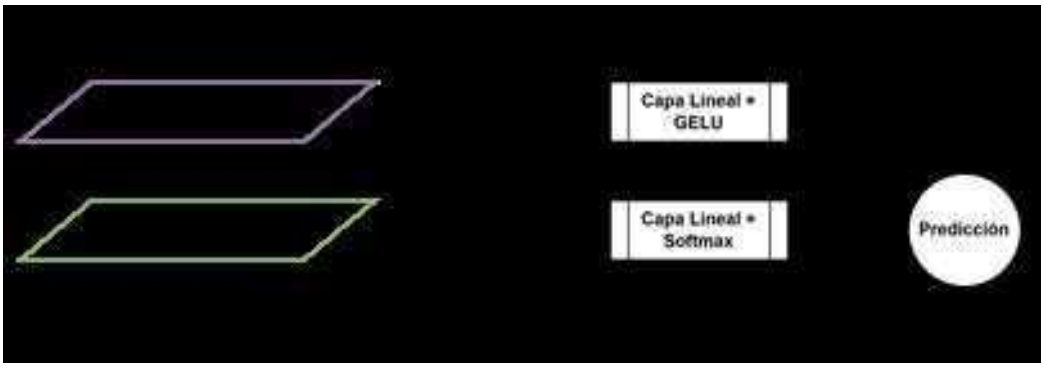

Figura 4.6: Clasificador de la Fusión de Características.

Capítulo 5
Experimentos

En este capítulo se describen la configuración experimental del método propuesto y resultados con su discusión. Se incluye una descripción del conjunto de datos, la configuración de los modelos complementarios y la búsqueda de hiperparámetros.

5.1. Conjunto de Datos

El conjunto de datos empleado en este trabajo de tesis corresponde a un subconjunto de Pitt Corpus (Becker et al. 1994), denominado ADReSSo (Luz, Haider, De La Fuente et al. 2021). Este conjunto de datos, compuesto por audios, cuenta con 237 instancias separadas en 166 de entrenamiento y 71 de prueba. Está separado en dos clases: Alzheimer (AD) y Grupo de Control (HC). A diferencia de ADReSS y Pitt Corpus, ADReSSo no cuenta con transcripciones manuales. Estadísticas de los audios se muestran en la tabla 5.1. Al analizar la puntuación del MMSE, se observa cierto solapamiento en el rango superior. Aunque los autores del conjunto de datos no ofrecen una explicación específica, se puede plantear una explicación sencilla: a pesar de obtener puntuaciones relativamente altas, el diagnóstico de demencia se realizó mediante pruebas de laboratorio e imágenes, en lugar de basarse únicamente en el MMSE.

Clase	#	Rango MMSE	Longitud Promedio
Alzheimer (AD)	122	[3-30]	85.26s
Grupo de Control (HC)	115	[24-30]	68.01s

Tabla 5.1: Distribución original de los datos.

Se empleó el criterio propuesto por Zaudig (1992) para establecer categorías basadas en el MMSE. Según este criterio, se establece un punto de corte de 27 para distinguir entre

individuos del grupo de control y aquellos con Alzheimer. Dentro del conjunto de individuos con Alzheimer (AD), hay una subcategoría específica que engloba a los pacientes con deterioro cognitivo leve (MCI). Para ellos, se ha definido un rango de 23 a 27. Sin embargo, es esencial señalar que la identificación se maneja como una tarea de clasificación binaria. En este proceso, el clasificador únicamente asigna una etiqueta: AD o HC. La categoría de deterioro cognitivo leve es de bastante interés, debido a que este grupo presenta casos de alta complejidad. Esto es debido al estar en las fases iniciales del Alzheimer, su diferenciación respecto a HC no es siempre evidente.

Clase	#	Rango MMSE
Alzheimer (AD)	95	$\leq$ 27
Grupo de Control (HC)	105	>27
D. Cognitivo Leve (MCI)	37	[23-27]

Tabla 5.2: Criterio de valores MMSE por cada clase.

5.2. Preprocesamiento y Selección de Características

A continuación se describe los diferentes procedimientos para obtener características de audio y transcripciones. Se describen cómo se obtuvieron las transcripciones y pausas, además, cómo se realizó la selección de características, el *fine-tuning* de BERT y la búsqueda de hiperparámetros.

5.2.1. Reconocimiento Automático de Voz

El conjunto de ADReSSo no tiene transcripciones manuales. Las características extraídas de texto son relevantes, particularmente en el inglés (Pérez-Toro et al. 2022), por lo tanto, para obtenerlas es necesario utilizar un modelo de ASR.

En primera instancia se utilizó *wav2vec 2.0* (Baevski et al. 2020), un modelo desarrollado por Facebook. Sin embargo, se observó que presentaba un elevado número de errores. Revisando de forma manual, las transcripciones presentaban una gran cantidad de omisiones. En el estado del arte se mencionó que este tipo de error, en particular, tiene el mayor impacto en los clasificadores. Por lo tanto, fue necesario reemplazar el modelo por uno que tuviera mayor resistencia al ruido.

Del estado del arte se decidió utilizar Whisper, un modelo ASR desarrollado por OpenAI. Este modelo fue entrenado con 680,000 horas de audio plurilingüe (Radford et al. 2023), lo que supone una ventaja sobre el resto de modelos ASR disponibles. Además, uno de sus puntos a destacar, es su resistencia al ruido. Para verificar esto, se realizó una sencilla prueba que consistió en transcribir canciones, y observar si el modelo podía distinguir la letra de la música. Al analizar los resultados, se observó que el modelo pudo inferir correctamente las letras a pesar del ruido de las percusiones. En otros modelos, cualquier ruido podría ser interpretado erróneamente y transcribirse como un sonido inexistente. Debido a estos resultados, se decidió utilizar Whisper. Este modelo viene preentrenado en diferentes tamaños y se eligió el modelo *large* para este caso. Al revisar las transcripciones generadas, no se encontraron omisiones de palabras relevantes. Además, cabe destacar que el modelo intenta transcribir algunas muletillas como "hmm" o "mhm".

Se ajustó la tasa de muestreo de los audios para adaptarse a los requisitos de cada modelo ASR, ya que cada uno necesita una tasa específica. En muchos casos, ésta suele ser baja, como 8 kHz o 16 kHz. Los audios de ADReSSo se grabaron originalmente a 44.1 kHz, por lo tanto, se convirtieron a una tasa de 16 kHz. En la figura 5.1, se muestra un histograma de los tókenes por transcripción, en el cual se pueden observar las diferencias que existen entre el grupo de AD y HC. En promedio, ambos grupos son similares con 151.48 (AD) y 151.37 (HC) tókenes.

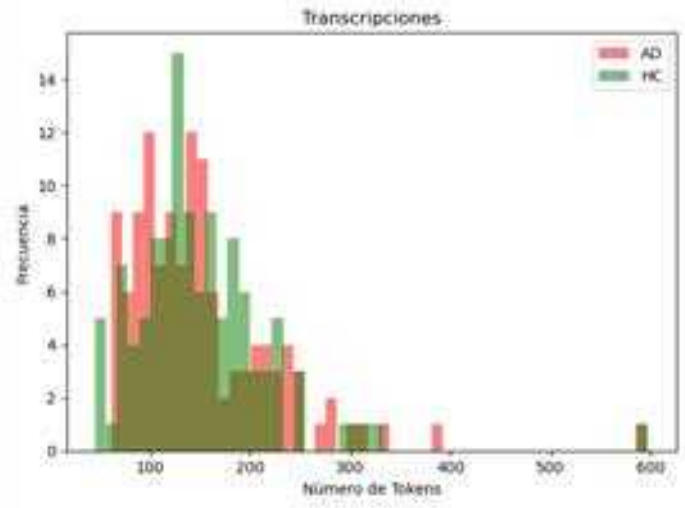

Figura 5.1: Histograma de Longitudes de Transcripciones.

Por otro lado, en la figura 5.2 se comparan las longitudes de los audios, en donde, se observa una proporción inversa. Esto permite observar a simple vista la presencia de pausas,

en donde los audios de los AD son más largos, pero al mismo tiempo las transcripciones son más cortas.

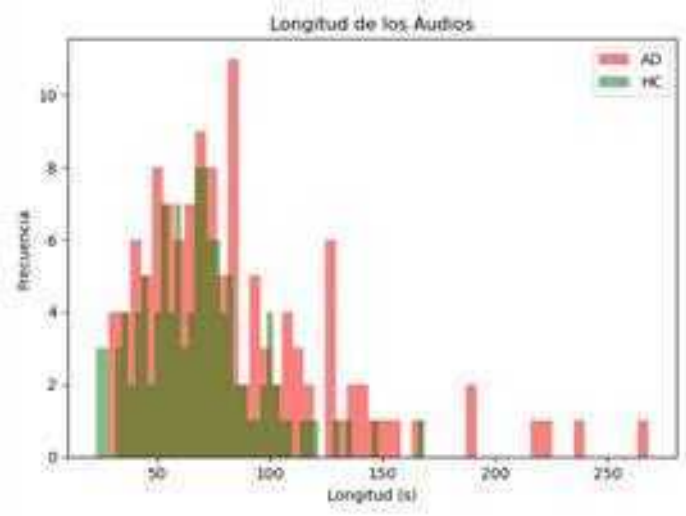

Figura 5.2: Histograma de Longitudes de los Audios.

5.2.2. Detección de Voz y Pausas

Para la detección de pausas se utilizó WebRTC (*WebRTC: Real-Time Communication in Browsers* s.f.), desarrollado por Google. WebRTC es un proyecto de código abierto que facilita la comunicación en tiempo real en navegadores web y aplicaciones móviles a través de una *application programming interfaces* (API). Esta herramienta incluye un modelo preentrenado detector de voz que clasifica segmentos de un mismo audio. Para su funcionamiento se preprocesaron los audios como se describió en la sección 5.2.1. Los audios se segmentaron en ventanas de 30 milisegundos. La salida del modelo es una secuencia binaria, donde un silencio es un 0 y voz es un 1. Una secuencia de 0 continuos se considera una sola pausa. Para describir su posición se guardó la posición del primer segmento con 0. Para obtener la duración de la pausa se cuenta el número de segmentos y se multiplica por el tamaño de la ventana, en este caso 30ms, por ejemplo, si una pausa se compone de 17 segmentos, su duración es 0.51s. Se calcula la duración de cada pausa para construir un histograma de pausas con los intervalos: [0.5s, 1s), [1s, 2s), [2s, 5s) y [5s, ∞). Y con las posiciones de las pausas se obtiene la dispersión con la fórmula 4.2.1. Se construye un segundo histograma de pausas utilizando la dispersión de cada intervalo para observar la distribución en cada uno.

5.2.3. Detección de Disfluencias

La detección de disfluencias se realiza mediante un modelo preentrenado propuesto por Lou et al. (2020). Este modelo utiliza un parser basado en Transformers para detectar los constituyentes que comprenden el reparandum y reparación de la disfluencia. Este modelo es capaz de detectar repetición de palabras, algunas muletillas, y autocorrecciones, es decir, situaciones en donde el hablante dice una frase e inmediatamente procede a corregirse con una frase similar. Existen varias versiones del modelo que utiliza diferentes tipos de embedding. La variante utilizada fue el modelo que utiliza embbeddings de BERT. En cuanto su utilización directa, el modelo no tiene hiperparámetros y solo requiere del texto de entrada. La salida es la transcripción en donde se coloca un token especial a lado de cada token. Si el token especial es E, indica que pertenece a una probable disfluencia, si el token es un guión bajo la palabra es correcta. A continuación se muestra un ejemplo:

$$\ldots \text{ the } \_ \text{ top } \_ \text{ of } \_ \text{ a } \underline{E} \text{ ladder } \underline{E} \text{ uh } \underline{E} \text{ a } \_ \text{ stool } \_ \text{ reaching } \_ \text{ for } \_ \ldots$$

Se realiza un conteo de cuantos tókenes conforman cada disfluencia para construir un histograma de tres intervalos: [1, 3), [3, 5), [5, ∞). Además, se obtiene la posición para calcular la dispersión de las disfluencias utilizando la fórmula 4.2.1.

5.2.4. Selección de Características

En los primeros intentos por ajustar la SVM al conjunto de estilo se obtuvieron resultados pobres, lo que planteó sospechas sobre la presencia de características poco informativas. Se observó que algunas categorías gramaticales mostraban valores constantes de 0, lo que indica que el parser de partes de la oración no detectaba ninguna palabra perteneciente a esas categorías. Para evitar la introducción de características poco informativas que generen ruido, se llevó a cabo una selección de características utilizando el método *one-way F-ANOVA*. Esta prueba se basa en el estadístico F, que en este contexto representa la relación entre la variación de las medias entre las clases y la variación dentro de las muestras.

Se calcula el valor de F para cada característica, y cuanto mayor sea este valor, mayor será la diferencia en términos de media entre las diversas clases. Por lo tanto, se supone que la característica es más informativa al describir de manera distinta a ambas clases. Por

el contrario, si los valores de la característica son similares para ambas clases, se considera ambigua y no aporta información relevante. Se calcula el p-valor al obtener el complemento de la función de distribución acumulada de F. En donde la hipótesis nula que corresponde a que la media de una característica en ambas clases es igual. Se filtraron las características con un p-valor mayor a 0.05. Como resultado, el conjunto de características se redujo de 77 a 31.

5.2.5. *Fine-Tuning* de BERT

BERT es uno de los modelos más utilizados en el estado del arte en la tarea de detección Alzheimer, lo que permite una mayor comparabilidad con otros trabajos. En términos generales, BERT presenta buenos resultados en la detección de Alzheimer. En M. S. S. Syed et al. (2020), se experimentó con la versión base y large del modelo. Los resultados indicaron que no había una mejora aparente en los resultados, por lo que para efectos de este trabajo se utilizó la versión base, que es menos costosa de realizar el *fine-tuning*. Se realizaron pruebas con BERT sin realizar *fine-tuning*, lo cual resultó en clasificación aleatoria. Por lo tanto, se decidió realizar *fine-tuning* utilizando el conjunto de entrenamiento.

Se probaron diversas configuraciones de épocas, tazas de aprendizaje y regularizadores de AdamW. Se utilizó *5-fold cross validation* y se buscó maximizar el valor de F1 promedio de las particiones.

5.2.6. Búsqueda de Hiperparámetros

Se utilizó un algoritmo heurístico denominado *Tree-structured Parzen Estimator* (TPE) (Bergstra et al. 2011) para la búsqueda de hiperparámetros. En términos simples, TPE es un método para optimizar funciones de caja negra que puede ser utilizado para encontrar los hiperparámetros de clasificadores. La idea central detrás de TPE, es dividir espacio de hiperparámetros de un modelo en dos zonas, una de candidatos buenos y otra de candidatos malos. Entonces, a partir de resultados iniciales con la función de caja negra, se modelan dos distribuciones de probabilidad distintas, $l(x)$ para los candidatos buenos y $g(x)$ para los malos. De forma iterativa, se extraen muestras de la distribución $l(x)$ y se evalúan a través de la función. Estos resultados son utilizados para adaptar y mejorar las distribuciones $l(x)$ y $g(x)$ respectivamente. El objetivo es estrechar la distribución $l(x)$ alrededor del mejor valor mientras se esparce $g(x)$. Este proceso continúa hasta que se cumpla una condición de parada o un número total de iteraciones.

Clasificador de Contenido y Fusión

Se utilizó un MLP para clasificar las características de contenido, de igual forma, para las representaciones obtenidas de la fusión temprana y GMU. En todos los casos, se probaron diferentes configuraciones de épocas, tasas de aprendizaje, *momentum* y *weight decay*. Como métrica de evaluación se utilizó el coeficiente de Matthews (Matthews 1975).

Los rangos de la búsqueda y los mejores valores para cada hiperparámetro se muestran en las tablas 5.3, 5.4 y 5.5.

Tabla 5.3: Hiperparámetros del clasificador de contenido (MLP).

Hyperarámetro	Rango de Búsqueda	Mejor Valor
Épocas	[1, 32]	15
Tasa de Aprendizaje	$[1 \times 10^{-8}, 0.1]$	0.0109
Momentum	[0, 0.99]	0.3288
Weight Decay	[0.0001, 0.1]	0.0007

Tabla 5.4: Hiperparámetros del clasificador de la fusión temprana.

Hyperarámetro	Rango de Búsqueda	Mejor Valor
Épocas	[1, 32]	23
Tasa de Aprendizaje	$[1 \times 10^{-8}, 0.1]$	0.0181
Momentum	[0, 0.99]	0.1628
Weight Decay	[0.0001, 0.1]	0.0013

Tabla 5.5: Hiperparámetros del clasificador de la fusión con GMU.

Hyperarámetro	Rango de Búsqueda	Mejor Valor
Épocas	[1, 32]	23
Tasa de Aprendizaje	$[1 \times 10^{-8}, 0.1]$	0.0522
Momentum	[0, 0.99]	0.1716
Weight Decay	[0.0001, 0.1]	0.0027

Clasificador de Estilo

El clasificador de estilo es un SVM, por lo tanto, los hiperparámetros a consideración fueron: el tipo de kernel, regularizador C, y en donde aplicara, grado de polinomio y término independiente. Para el caso de hiperparámetro gamma, este se determinó mediante una fórmula: $1/n$, donde n es el número de características. De igual forma, como métrica de evaluación se utilizó el coeficiente de Matthews.

Al finalizar la búsqueda, se determinó que la mejor configuración utiliza un kernel sigmoide, en consecuencia, el valor del grado de polinomio es omitido. Los rangos de la búsqueda y los mejores valores para cada hiperparámetro se muestran en la tabla 5.6.

Tabla 5.6: Hiperparámetros del clasificador de estilo (SVM).

Hyperarámetro	Rango de Búsqueda	Mejor Valor
Kernel	Sigmoide, Lineal, Radial, Polinomial	Sigmoide
Gamma	-	0.0323
Regularizador C	[0.001, 5]	4.6063
Término Independiente	[0.01, 3]	1.2152
Grado de polinomio	[1, 10]	-

5.2.7. Métricas de Evaluación

A continuación, se detallan las métricas de evaluación (Powers et al. 2020) empleadas en los experimentos. Las fórmulas se expresan en términos de clases positivas y negativas, donde P representa la clase positiva y N la clase negativa. Adicionalmente, se definen TP y FP como verdaderos positivos y falsos positivos respectivamente. Mientras que TN y FN corresponden a verdaderos negativos y falsos negativos. Cabe mencionar que los estudios del estado del arte utilizan principalmente a la exactitud como métrica de referencia, mientras que el coeficiente de correlación de Matthews (MCC) (Matthews 1975) fue utilizado en la optimización de hiperparámetros.

Exactitud

Es el porcentaje de instancias clasificadas correctamente.

$$\frac{TP + TN}{P + N}$$

Precisión

El porcentaje que corresponde a la proporción de verdaderos positivos en todas las predicciones de clase positiva. También, puede interpretarse como la probabilidad de que una predicción positiva sea realmente positiva.

$$\frac{TP}{TP + FP}$$

Recuerdo

El porcentaje que corresponde a la proporción de verdaderos positivos de toda la población de clase positiva del conjunto a evaluar. En otras palabras, equivale a la probabilidad de que el modelo clasifique una instancia como positiva.

$$\frac{TP}{TP + FN}$$

F1

Es la media armónica de la precisión y recuerdo, permitiendo generalizar ambas en una sola métrica. Esto alivia problemas relacionados con la precisión y recuerdo.

$$\frac{2TP}{2TP + FP + FN}$$

Coeficiente de Correlación de Matthews (MCC)

Esta métrica mide el nivel de asociación entre dos variables. A diferencia de las otras métricas, MCC tiene un rango que va de -1 a 1, siendo 1 una clasificación perfecta.

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

5.3. Resultados

En esta sección, se presentan los resultados de cada modalidad, así cómo su respectiva discusión. También se introducen distintas estrategias para combinar sus resultados.

5.3.1. Estilo vs. Contenido

A continuación, se presentan los resultados del clasificador de estilo y contenido de forma individual. Se utilizaron las 71 instancias de prueba para obtener estos resultados. Según en las métricas de clasificación se muestra que ambos métodos obtienen buenos resultados en términos generales. Sin embargo, se observa que el clasificador de estilo muestra resultados más equilibrados, mientras que el clasificador de contenido presenta una precisión deficiente en el grupo de control (HC) y un recuerdo pobre en la clase de Alzheimer. Estas diferencias plantean la pregunta de si los clasificadores tienen un comportamiento similar o lo suficientemente diferente como para complementarse entre sí.

Tabla 5.7: Resultados del clasificador de contenido sobre el conjunto de prueba.

Clase	Precisión	Recuerdo	F1	Soporte
AD	0.96	0.71	0.82	35
HC	0.78	0.97	0.86	36
Macro Average	0.87	0.84	0.84	71
Exactitud			0.85	71
MCC			0.71	71

Tabla 5.8: Resultados del clasificador de estilo sobre el conjunto de prueba.

Clase	Precisión	Recuerdo	F1	Soporte
AD	0.86	0.86	0.86	35
HC	0.86	0.86	0.86	36
Macro Average	0.86	0.86	0.86	71
Exactitud			0.86	71
MCC			0.72	71

En las tablas 5.7 y 5.8, se observa que ambos métodos tienen una exactitud similar. Por lo tanto, es importante determinar si existe una complementariedad entre ambos o si son equivalentes. Para verificarlo se aplicó la fórmula de *Coincident Failure Diversity* (CFD) (W. Wang 2008). Esta fórmula permite conocer si un conjunto de clasificadores presentan

los mismos errores, o en otras palabras, si existe una diversidad en los errores. CFD se define como:

$$\text{CFD} = \begin{cases} \frac{1}{1-p_0} \sum_{r=1}^{N} \frac{N-r}{N-1} p_r, & Si\ p_0 < 1 \\ 0, & Si\ p_0 = 1 \end{cases} \tag{5.3.1}$$

Donde Pn es la probabilidad de que n clasificadores fallen al mismo tiempo y está definida como kn/M, siendo kn el número de instancias fallidas en los n clasificadores y M, el total de instancias. N corresponde al número total de clasificadores. Si $p0$ equivale a 1, implica que la probabilidad de fallas en 0 clasificadores es del 100 %, lo que señala que los clasificadores no comenten errores. Un valor cercano a 1 indica que los errores son únicos, mientras que cercano a 0 indica que comete los mismos errores y no hay diferencias entre los métodos.

Al aplicar la fórmula a los resultados obtenidos se obtuvo un CFD: 0.83. Lo que indica que existen diferencias notables entre los errores de los métodos de contenido y estilo. Por consiguiente, se analizó el comportamiento de los clasificadores en 3 rangos diferentes de MMSE, considerando los intervalos: 0-22 AD, 23-27 MCI y 28-30 HC. En la tabla 5.9, se muestran los resultados en cada intervalo en términos de la exactitud, es decir, se verificó si los métodos clasificaban correctamente las instancias en ese rango. Para el caso de los MCI se espera que sean clasificados como AD, debido a que si bien su deterioro cognitivo no es grave, está fuera de rangos normales. Se puede observar una diferencia notable en como los métodos detectan a los casos de MCI, siendo el clasificador de estilo el que obtiene una exactitud superior. Esto indica que el método de contenido no puede diferenciar correctamente a un paciente en una etapa inicial de uno completamente sano, por lo tanto, su clasificación es prácticamente aleatoria. La explicación de esto es que, aunque en las etapas iniciales no se manifiestan cambios sustanciales que modifiquen la semántica del diálogo, ya se empiezan a percibir cambios en la manera de expresarse.

5.3.2. Fusión de Características

A continuación se presentan los resultados sobre el conjunto de prueba. Los resultados incluyen puntajes de precisión, recuerdo, F1 y sus respectivos promedios ponderados. Para facilitar la comparación, también se proporcionan los valores de exactitud y coeficiente de correlación de Matthews. Este último valor es un número entre -1 y 1 que generaliza la matriz de confusión y, a diferencia de F1, tiene en cuenta los negativos verdaderos. Un

Tabla 5.9: Resultados del clasificador de estilo sobre el conjunto de prueba.

Método	Clase	Exactitud
Estilo (SVM)	Alzheimer (AD)	0.83
	Deterioro Cognitivo Leve (MCI)	0.93
	Grupo de Control (HC)	0.85
Contenido (MLP)	Alzheimer (AD)	0.79
	Deterioro Cognitivo Leve (MCI)	0.57
	Grupo de Control (HC)	1.00

valor de 1 indica una clasificación perfecta, -1 indica una clasificación invertida y un valor cercano a 0 indica una clasificación aleatoria.

Tabla 5.10: Resultados de fusión temprana.

Clase	Precisión	Recuerdo	F1	Soporte
AD	0.96	0.77	0.86	35
HC	0.81	0.97	0.89	36
Macro Average	0.89	0.87	0.87	71
Exactitud			0.87	71
MCC			0.76	71

Los resultados obtenidos son similares entre ellos. En el caso de la fusión temprana y con centroides, su diferencia radica que en este último presenta un mayor recuerdo. Los resultados de la fusión con GMU son únicamente superiores a los resultados obtenidos por el clasificador de contenido, indicando que su capacidad de ponderar entre ambas modalidades es limitada en este caso, o por lo menos no se traduce en resultados superiores a una fusión temprana simple. Todos los clasificadores ven un aumento en su capacidad de clasificar instancias del grupo de control (HC) debido a la integración de características de contenido.

Los resultados presentados en las tablas 5.10 y 5.12 guardan similitudes, aunque la principal distinción radica en cómo se distribuyen la precisión y el recuerdo entre el método que emplea fusión y el que utiliza centroides. Un recuerdo elevado muestra la habilidad

Tabla 5.11: Resultados de fusión con GMU.

Clase	Precisión	Recuerdo	F1	Soporte
AD	0.96	0.74	0.84	35
HC	0.80	0.97	0.88	36
Macro Average	0.88	0.86	0.86	71
Exactitud			0.86	71
MCC			0.74	71

Tabla 5.12: Resultados de fusión mediante centroides.

Clase	Precisión	Recuerdo	F1	Soporte
AD	0.91	0.83	0.87	35
HC	0.85	0.92	0.88	36
Macro Average	0.88	0.87	0.87	71
Exactitud			0.87	71
MCC			0.75	71

del método para identificar los casos de Alzheimer dentro de la población. En contraposición, una precisión alta significa que los casos detectados de Alzheimer corresponden realmente a esta enfermedad. Por consiguiente, la elección del enfoque más adecuado se basa en lo que se quiera priorizar. Si se considera crítico minimizar la cantidad de falsos negativos (es decir, individuos con Alzheimer clasificados erróneamente como saludables) se debería optar por un método con un recuerdo alto, en este caso el método de selección de clasificadores. Sin embargo, si el coste asociado a los falsos positivos (personas sanas identificadas erróneamente como casos de Alzheimer) es mayor, entonces la opción recomendable sería el método que ofrece una alta precisión, en este caso el método de fusión. No obstante, si se asume que tanto los falsos positivos como los falsos negativos tienen un coste equivalente, el método de fusión presenta la ventaja de ser el más sencillo, y por ende, más fácil de implementar y entrenar. En cuanto a los resultados obtenidos con la GMU, aunque no superan a los de los otros dos métodos, estos sí logran superar al baseline.

5.3.3. Discusión

En la tabla 5.13 se muestran la comparación de resultados entre los métodos propuestos y el baseline. A partir de los resultados obtenidos en los experimentos por modalidad, donde se llevó a cabo la clasificación utilizando exclusivamente un conjunto de características, se observa que el clasificador de estilo con el conjunto de características seleccionado, alcanza resultados competitivos en comparación con BERT, el cual también emplea las mismas transcripciones para su fine-tuning. El clasificador de contenido supera a BERT, lo que sugiere que la nueva representación ofrece una descripción que logra una separación mejor entre las clases. Las convoluciones de la CNN convierten los embeddings en una representación que ilustra las relaciones de dependencia entre los tókenes. Lo que podría significar que la pérdida de vocabulario y la simplificación de frases logran la formación de patrones en términos de estas relaciones que mejoran la clasificación.

Tabla 5.13: Comparación con *baselines*.

Método / Modelo	Exactitud	F1	MCC
MLP (Contenido)	0.85	0.84	0.71
SVM (Estilo)	0.86	0.86	0.72
Fusión Temprana	**0.87**	**0.87**	**0.76**
Fusión con GMU	0.86	0.86	0.74
Fusión mediante Selección de Clasificadores	0.87	0.87	0.75
BOW	0.77	0.77	0.55
BOW (2-gram)	0.82	0.82	0.63
BOW (3-gram)	0.82	0.82	0.63
BOW (4-gram)	0.80	0.80	0.60
BOW (5-gram)	0.75	0.73	0.47
BERT	0.83	0.83	0.67

Tras evaluar el rendimiento de los clasificadores de Contenido y Estilo para cada categoría: Alzheimer (AD), Deterioro Cognitivo Leve (MCI) y Grupo de Control (HC), se observa que en la clasificación de AD, el estilo posee una ventaja ligera sobre el contenido con un margen del 4 %. En términos generales, ambos resultados son parecidos. No obstante, al analizar el desempeño en el Grupo de Control, el contenido supera al estilo por

un 15 %, identificando con exactitud todas las instancias. Esto implica que el componente semántico es crucial para reconocer individuos con cognición intacta. Es importante mencionar que una alta exactitud en HC, pero un menor rendimiento en las otras dos categorías, implica que el clasificador confunde varias instancias en particular las de MCI. Los casos clasificados como MCI se ubican en la frontera entre AD y HC, siendo su posición en la escala de MMSE próxima a HC. Esta cercanía puede explicar el pobre resultado, ya que desde una perspectiva semántica, el MCI no muestra patrones distintivos que faciliten una diferenciación con HC. Lo que significa que: no experimentan una degradación del lenguaje lo suficientemente notoria como para ser detectada en los embeddings. Por su parte, el clasificador de estilo demuestra ser eficiente al identificar estos casos limítrofes que conforman el MCI. Por lo tanto, esto indica que para la detección de MCI, los aspectos estructurales del diálogo son más determinantes que el propio contenido semántico de las palabras.

Al examinar la complementariedad entre contenido y estilo mediante CFD, se registró un coeficiente de 0.83. Un valor próximo a 1 indica una complementariedad alta, lo cual motivó a explorar la integración de ambos conjuntos de características. Se implementaron tres técnicas: fusión temprana, fusión mediante GMU y fusión mediante selección de clasificadores. Todas obtuvieron resultados similares, siendo la fusión temprana levemente superior. De las tres, la fusión mediante GMU resultó ser la menos eficiente, posiblemente debido a una falta de datos, dificultando el entrenamiento de las unidades. Respecto a la fusión mediante selección de clasificadores, aunque supera a la fusión temprana en la recuperación de la clase AD, muestra una leve reducción en la precisión. Aunque en términos generales, la fusión anticipada es superior considerando que, pese a su simplicidad, obtiene los mejores resultados.

Al contrastar los métodos de fusión con modelos más simples como el *bag of words*, es claro que se alcanzan resultados superiores. Es importante destacar, que el enfoque que solamente considera características estilísticas arroja mejores resultados, además, tiene la ventaja de ser insensible a la temática del texto proporcionando cierta resistencia ante palabras que no pertenecen al vocabulario.

5.3.4. Comparación con el Estado del Arte

Como puntos de referencia para comparar el rendimiento del método propuesto contra un método simple, se realizó una clasificación mediante *bag of words* (BOW) extraídas de las transcripciones. Se consideraron de 1 a 5-*grams* para construir representaciones que se basan en segmentos continuos de tókenes. Como clasificador en todos los casos se utilizó SVM, y fue entrenado de la misma forma que se especifica en la sección 4.2.3.

Para medir si existe una mejora en el caso de las características de contenido, es necesario comparar con BERT en su implementación original. Por lo tanto, sobre un BERT *base* se realizó *fine-tuning* con los mismos hiperparámetros encontrados mediante el método especificado en la sección 5.2.5 con el objetivo de facilitar su comparación.

Los trabajos del estado del arte (Z. S. Syed et al. 2021; Pan, Mirheidari, Harris et al. 2021; Rohanian et al. 2021; Pappagari et al. 2021; Chen et al. 2021) se caracterizan por utilizar el mismo conjunto de datos de ADReSSo 2021 (Luz, Haider, De La Fuente et al. 2021), por lo tanto, todos los trabajos utilizan transcripciones automáticas obtenidas mediante algún modelo de ASR. Se omitieron artículos que trabajan solo con transcripciones manuales, debido a que estas carecen de los errores comúnmente cometidos por los sistemas de ASR, lo cual no permite una comparación adecuada.

En relación con las representaciones empleadas en los trabajos del estado del arte, la mayoría de los casos se recurrieron a características lingüísticas, aunque con diferencias en qué tipos de características fueron finalmente empleadas. Se omitió el uso de características acústicas, dado que el presente trabajo se centra en analizar la parte estilística del lenguaje de personas afectadas por Alzheimer. En la tabla 5.14 se describen los puntos importantes de cada trabajo del estado del arte.

Tabla 5.14: Resumen de trabajos del estado del arte.

Método / Modelo	Descripción
Método Propuesto	▪ Se propone un método que fusiona dos representaciones, una de contenido que abstrae la parte semántica de transcripciones mediante una red CNN que aprender relaciones entre tókenes, representados por embeddings de BERT, y otra de estilo que se compone de características como categorías gramaticales, profundidad de árboles sintácticos, diversidad de vocabulario, histogramas de pausas y disfluencias. ▪ Se emplean medidas de dispersión de pausas y disfluencias para determinar su prevalencia en el habla de un examinado. ▪ Para la clasificación, ambas representaciones son concatenadas y enviadas a un perceptrón multicapa que realiza la predicción.
Z. S. Syed et al. 2021	▪ Se emplearon diferentes técnicas de preprocesamiento de transcripciones para posteriormente generar embeddings contextualizados de cada token con Facebook BART; Para obtener una sola representación, se utilizó diferentes capas de agrupamiento con los embeddings de cada token. ▪ Se compararon contra características lingüística como histogramas de categorías gramaticales y dependencia gramatical, medidas de legibilidad y diversidad de vocabulario. ▪ Ambas representaciones fueron clasificadas mediante SVM y LR.

Continúa en la siguiente página

Tabla 5.14: Resumen de trabajos del estado del arte. (Continuación)

Pan, Mirheidari, Harris et al. 2021	▪ Para la representación, concatenaron embeddings acústicos de wav2vec 2.0 con embeddings de BERT. ▪ Se realizó aumento de datos mediante hipótesis de ASR de wav2vec 2.0. ▪ Un MLP fue utilizado como clasificador.
Rohanian et al. 2021	▪ Se entrenó una LSTM con una arquitectura de compuertas, en diferentes tipos de características que incluían: embeddings de GloVe, características acústicas COVAREP, probabilidad de palabras, disfluencias y pausas.
Pappagari et al. 2021	▪ Se utilizó BERT ajustado a transcripciones obtenidas de un servicio de transcripción automática de *Amazon Web Services*(AWS).
Chen et al. 2021	▪ Se propuso un ensamble de modelos de regresión logística, utilizando una votación basada en el promedio de la probabilidad de cada clase, en cada predicción. ▪ Como entrada se utilizaron diferentes combinaciones de embeddings de BERT (tókenes y oraciones) y Linguistic Inquiry and Word Count (LIWC) con características acústicas como eGeMAPS, ComParE y IS10-Paraling; con el objetivo de explotar la complementariedad entre modelos.

En la tabla 5.15 se muestran los resultados tanto de los métodos propuestos como del estado del arte. En términos generales, las mejores soluciones del estado del arte reportan una exactitud del 0.85, mientras que los métodos de fusión se ubican en un 2 % por arriba, al alcanzar una exactitud de 0.87. Esto sugiere que la estrategia de emplear dos modalidades es efectiva, ya que consigue identificar algunos casos complicados dentro del conjunto de MCI, lo que finalmente aumenta el desempeño de los métodos en general.

Tabla 5.15: Comparación con el estado del arte.

Método / Modelo	Exactitud	F1	MCC
MLP (Contenido)	0.85	0.84	0.71
SVM (Estilo)	0.86	0.86	0.72
Fusión Temprana	**0.87**	**0.87**	**0.76**
Fusión con GMU	0.86	0.86	0.74
Fusión mediante Selección de Clasificadores	0.87	0.87	0.75
BOW	0.77	0.77	0.55
BOW (2-gram)	0.82	0.82	0.63
BOW (3-gram)	0.82	0.82	0.63
BOW (4-gram)	0.80	0.80	0.60
BOW (5-gram)	0.75	0.73	0.47
BERT	0.83	0.83	0.67
Z. S. Syed et al. 2021	0.85	-	-
Pan, Mirheidari, Harris et al. 2021	0.85	-	-
Rohanian et al. 2021	0.84	-	-
Pappagari et al. 2021	0.85	-	-
Chen et al. 2021	0.82	-	-

Capítulo 6

Análisis de Resultados

En esta sección se discuten los resultados obtenidos, y se presentan los análisis acerca de sus características, para indagar más acerca de como influyen en el MMSE y su relevancia. Además, se analizan instancias erróneas para detectar problemas relacionados con el uso de los modelos de contenido y estilo por separado.

6.1. Estudio de Ablación

Para medir el impacto de ciertos conjuntos de características, se realizó un estudio de ablación utilizando las características de estilo. La división de estos conjuntos se muestra en la tabla 6.1. Para el estudio se entrenó un SVM excluyendo un conjunto de características a la vez y utilizando los mismos hiperparámetros encontrados durante el entrenamiento original descrito en la sección 4.2.3.

Los resultados se muestran en la tabla 6.2, en donde se puede observar que las partes de la oración (que describen la composición del diálogo) tienen una relevancia importante, dado que se observa una reducción del 21 % con eliminación, aunque cabe señalar que es el grupo más numeroso al contar con 15 características.

En el caso de las pausas hay una reducción del 15 % y se observa que su eliminación afecta en particular la detección en los casos del grupo de control. Por otro lado, la eliminación de las disfluencias conlleva una reducción del 15 %, afectando más a la detección de los casos de Alzheimer. Finalmente, el conjunto de complejidad que solo está compuesto por dos características tiene el menor impacto con una reducción de sólo 3 % y no se aprecia alguna afectación particular en las clases.

Tabla 6.1: Conjuntos de características.

Categoría	Características	#
Partes de la oración	Signos de Puntuación (... / . / ?), Puntos Suspensivos, Número Cardinal, Determinante, Existencial "There", Conjunción Subordinada, Adjetivo, Sustantivo (Singular), Pronombre Personal, Pronombre Posesivo, Adverbio, Adverbio de Partícula, Interjección, Verbos (3ra persona, Singular), Pronombre Interrogativo Personal	15
Pausas	Pausa [1s, 2s), Pausa [2s, 5s), Cantidad de Pausas, Pausa Mínima, Pausa Total, Dispersión de Pausa [0.5s, 1s), Dispersión de Pausa [1s, 2s), Dispersión de Pausa [2s, 5s), Dispersión de Pausa [5s, ∞).	9
Disfluencias	Disfluencia [1t, 3t), Disfluencia [3t, 5t), Disfluencia [5, ∞), Dispersión Global de Disfluencias, Dispersión de Disfluencia [1t, 3t)	5
Complejidad	Prof. Media del Árbol Sintáctico, Diversidad de Vocabulario	2

Tabla 6.2: Resultados de conjuntos excluidos.

Conjunto Excluido	F1 (AD)	F1 (HC)	F1	Exactitud	Reducción
Ninguno	0.86	0.86	0.86	0.86	
Partes de la oración	0.68	0.68	0.68	0.68	21 %
Pausas	0.75	0.72	0.73	0.73	15 %
Disfluencias	0.75	0.77	0.76	0.76	12 %
Complejidad	0.83	0.83	0.83	0.83	3 %

6.2. Correlación de Características de Estilo con MMSE

Como se ha mencionado previamente, el conjunto de datos de ADReSSo incluye una métrica que evalúa el estado cognitivo de los participantes denominada Mini Mental State Exam (MMSE). Esta métrica tiene un rango de 1 a 30, donde un valor bajo indica un problema serio en la cognición y un valor alto una cognición normal. Para comprender como las características de estilo se relacionan con el MMSE, se calculó el coeficiente de correlación de Pearson de cada una. En la tabla 6.3, se muestran las primeras 10 características con mayor coeficiente (r), tanto positivo como negativo. La lista exhaustiva se ubica en la tabla 7.2 que se ubica en el Anexo. Dentro de las correlaciones positivas, que

serán aquellas características cuyos valores altos referente a personas cognitivamente sanas, se puede observar que la característica que los sustantivos presentan una correlación moderada, lo que concuerda con lo descrito en literatura médica, que mencionan que la desaparición del lenguaje no es uniforme y suele afectar principalmente los sustantivos (Bird et al. 2000). Otras características que presentan también una correlación moderada son las conjunciones subordinadas, y la profundidad media de los árboles sintácticos, que describen la complejidad de las oraciones del participante. Como consecuencia de la pérdida de vocabulario y la disminución en la capacidad de analizar e interpretar, las descripciones de los pacientes con Alzheimer se vuelven más cortas. Determinantes (p.j., *the, a, some, that*) demostraron tener una correlación moderada, mientras que adjetivos, verbos en 3ra persona del singular y pronombres posesivos mostraron una correlación débil. Estas características indican que los participantes con Alzheimer suelen omitirlos al hablar, reflejando una simplificación del lenguaje. A pesar de su correlación débil, se ha observado que los números cardinales, es decir, las palabras que indican cantidades, podrían ser una característica que describa la pérdida de detalle en las descripciones de los participantes con Alzheimer. Estas palabras son más frecuentes en personas sanas, quienes poseen una mayor habilidad para comprender la composición de la escena y expresarse con mayor detalle.

Dentro de las características con una correlación negativa, se encuentran los pronombres interrogativos (p.j., “who”, “whom”, “what”, “whoever”, “whatever”), los cuales muestran una correlación moderada con un nivel bajo de MMSE. Esta correlación está asociada con la presencia de Alzheimer al reflejar la incertidumbre de los participantes al no comprender las instrucciones o lo que están observando, probablemente debido a dificultades cognitivas. También se observó un aumento en el uso de adverbios y pronombres, los cuales parecen utilizarse para reemplazar sustantivos en ciertas partes de las oraciones. Se calculó el coeficiente de Pearson entre sustantivos y adverbios (r: -0.5694), pronombres personales (r: -0.6392) y adverbios de partículas (r: -0.1605), y los resultados indicaron una fuerte correlación negativa. Dentro de las características con una correlación débil se encuentran los puntos suspensivos y otros signos de puntuación. Whisper tiene la capacidad de inferir interrupciones en el diálogo y las representa en las transcripciones como “...” en ciertas ocasiones. Estas interrupciones son más comunes entre los participantes con Alzheimer y reflejan dificultades para expresarse.

Además Whisper es capaz de inferir el final de una frase marcándolo con un “.”; o si dicha frase está formulada como una pregunta, la marca con un “?” al final. Un alto número

de estos signos indica la presencia de múltiples intervenciones cortas, interrupciones y dudas en el diálogo.

Las disfluencias, que se dividen en intervalos según el número de tókenes involucrados, también presentaron correlaciones moderadas. Específicamente, se encontró una correlación moderada entre las disfluencias compuestas por 1 a 3 tókenes y aquellas con más de 5 tókenes. Esto sugiere que las disfluencias son momentos en los que el participante se corrige a sí mismo. Finalmente, se encontró una correlación débil entre la dispersión de pausas de 1 a 2 segundos. Una alta dispersión de pausas indica la presencia de este fenómeno a lo largo del diálogo, algo que se observa en participantes con un MMSE bajo.

Tabla 6.3: Mayor correlación positiva y negativa de Pearson.

Correlación Positiva (+)		Correlación Negativa (-)	
Característica	r	Característica	r
Sustantivo (Singular)	0.4826	Pronombre Interrogativo Personal	-0.4881
Conjunción Subordinada	0.3633	Adverbio	-0.3744
Determinante	0.3425	Pronombre Personal	-0.3739
Prof. media de Árbol Sintáctico	0.3331	Interjección	-0.3597
Adjetivo	0.2851	Adverbio de Partícula	-0.3248
Sustantivo (Plural)	0.2325	Dispersión de Pausa [1s, 2s)	-0.286
Número Cardinal	0.2243	Puntos Suspensivos	-0.2775
Pronombre Posesivo	0.1922	Disfluencia [1t, 3t)	-0.2487
Diversidad de Vocabulario	0.1897	Signo de Puntuación (... / . / ?)	-0.2324
Verbo (3ra persona, Singular)	0.1881	Disfluencia [5t, ∞)	-0.2271

6.3. Relevancia de Características

Para cuantificar la capacidad descriptiva de las características a las instancias, se calculó su ganancia de información. La cual se obtiene de la diferencia entre la entropía de las clases y características. Si la diferencia es alta indica que las características son informativas, es decir, permiten una mejor separación entre las clases. En la tabla 6.4 se muestra las 10 características con mayor y menor ganancia de información, mientras que la lista completa se muestra en la tabla 7.3 localizada en el Anexo. Los resultados indican que las características más informativas son todas aquellas relacionadas con las pausas,

en particular la dispersión de pausas de [0.5s, 1s), la duración mínima y total. Esto indica que en el caso de la dispersión la diferencia entre la distribución de pausas en el diálogo es notable entre los participantes con Alzheimer y los del grupo de control. Por el otro, si la pausa más pequeña tiene cierta duración, es muy probable que sea debido a la falta de comprensión. Las medidas de complejidad: diversidad de vocabulario y profundidad media del árbol sintáctico se encontraron entre las 10 características con mayor ganancia. Ambas características describen la capacidad de los participantes para formular descripciones complejas y ricas en el uso de diferentes palabras. Los sustantivos y pronombres personales resultaron relevantes, y si toma en cuenta las correlaciones de la tabla 6.3, esto sugiere que existe una pérdida importante de sustantivos, mientras que el uso de pronombres personales aumenta entre los participantes de Alzheimer. Finalmente, como se mencionó previamente, los signos de puntuación señalan duda e interrupciones en el diálogo, la ganancia de información es alta debido a su mayor presencia en participantes con Alzheimer. En cuanto a las características con menor ganancia se encuentran en su mayoría aquellas relacionadas con las disfluencias. Es posible que el modelo de detección de disfluencias haya generado numerosos falsos positivos, lo que impactó en la estimación de la ganancia de información. Algo similar sucede con Whisper a la hora marcar interrupciones con puntos suspensivos. Es importante señalar que características como número cardinal, existencial *there* y pronombre "wh" obtuvieron ganancias de informacion superiores a las disfluencias.

Tabla 6.4: Resumen de mayor y menor ganancia de información.

Mayor Ganancia		Menor Ganancia	
Característica	IG(x|y)	Característica	IG(x|y)
Pausa Total	0.9983	Disfluencia [1t, 3t)	0.1008
Dispersion de Pausa [0.5s, 1s)	0.9983	Disfluencia [3t, 5t)	0.1193
Pausa Mínima	0.9983	Dispersión de Disfluencia [1t, 3t)	0.1635
Cantidad de Pausas	0.9863	Puntos Suspensivos	0.1702
Diversidad de Vocabulario	0.9622	Dispersión de Pausa [5s, ∞)	0.1738
Pausa [1s, 2s)	0.9513	Dispersión Global de Disfluencias	0.2053

Continúa en la siguiente página

Tabla 6.4: Resumen de mayor y menor ganancia de información. (Continuación)

Pronombre Personal	0.9410	Disfluencia [1t, 3t)	0.4340
Prof. media del Árbol Sintáctico	0.9260	Número Cardinal	0.4473
Signo de Puntuación	0.9185	Existencial *there*	0.5239
Sustantivo (Singular)	0.9163	Pronombre Interrogativo Personal	0.5465

6.4. Análisis de Errores

En esta sección se presentan el análisis cualitativo para determinar las causas detrás de ciertos errores, principalmente relacionados con la clase MCI, dado que es la clase de mayor interés y dificultad para clasificar, debido a que se ubican en la frontera entre las instancias de personas con Alzheimer y del Grupo de Control. En la Tabla 6.5, se observa la distribución de los errores en el conjunto de prueba con respecto a la clase MCI. La tabla está dividida en 3 categorías donde se muestra la cantidad neta de errores de cada clasificador, SVM para las características de estilo y MLP para el contenido como fue descrito en las secciones 4.2.3 y 4.1.3. De la distribución se puede ver que el clasificador de contenido tiene problemas para clasificar a la clase MCI, mientras que el clasificador de estilo en mayoría son errores de las otras clases. En cuanto a errores comunes, es decir, aquellas instancias que ningún modelo pudo clasificar correctamente se obtuvieron en total 3.

Tabla 6.5: Distribución de errores entre clasificadores.

Categoría	MCI	No MCI	Total
Contenido	6	5	11
Estilo	1	9	10
Común	1	2	3

Para analizar el desempeño de cada clasificador con respecto a cada clase, se obtuvo la exactitud de cada una. Los resultados se detallan en la Tabla 6.6. Las exactitudes indican que el modelo que utiliza características de estilo es mejor para detectar a las instancias de la clase de Alzheimer y Deterioro Cognitivo Leve (MCI), mientras que el clasificador de contenido es mejor para detectar las instancias del Grupo de Control. Al llevar a cabo la fusión, el método adquiere la habilidad de identificar casos de Alzheimer con la misma

exactitud del clasificador de estilo, y a los individuos del grupo de control con la exactitud del clasificador de contenido. Sin embargo, solo se observa un ligero incremento para la clase de MCI, lo que indica que las características de contenido entorpecen la capacidad del clasificador para diferenciar tal clase.

Tabla 6.6: Exactitudes por clase.

Modelo	Clase	Exactitud
Estilo (SVM)	Alzheimer (AD)	0.83
	Deterioro Cognitivo Leve (MCI)	0.93
	Grupo de Control	0.85
Contenido (MLP)	Alzheimer (AD)	0.79
	Deterioro Cognitivo Leve (MCI)	0.57
	Grupo de Control (HC)	1.00
Fusión (MLP)	Alzheimer (AD)	0.83
	Deterioro Cognitivo Leve (MCI)	0.64
	Grupo de Control (HC)	1.00

A continuación, se examinaron casos erróneos para identificar las razones por las que ni el clasificador de contenido ni el de estilo lograron etiquetar las instancias de manera adecuada. Se detectaron solamente 3 de estos casos, y se revisarán individualmente.

Caso 1

En el caso 1, se presenta una mujer de edad avanzada, la cual fue etiquetada como sana por los clasificadores, a pesar de estar diagnosticada con Alzheimer.

A continuación se presenta su transcripción:

> `Everything you see happening in that picture, everything that's going on in that picture. I see a boy stretching out for, I don't know, a bowl or a cake. And he was standing on a stool. His little sister was reaching out her hand and the ladder was beginning to go over. Mother was at the dishwasher and she left this faucet.`

Según su informe del MMSE, la participante obtiene una puntuación de 16 puntos, lo cual indica la presencia de un daño cognitivo grave. Al analizar el audio, se descubrió

que su duración es de 28 segundos, muy por debajo del promedio de 76 segundos. De hecho, resulta difícil determinar si el audio está completo, ya que finaliza de forma abrupta. El discurso de la participante carece de pausas significativas y, desde un punto de vista gramatical, presenta un error semántico al mencionar *ladder* (escalera) en lugar de *stool* (banquito). También se observa un uso extraño del verbo frasal *to go over* y la omisión del artículo antes de *Mother*. Por lo tanto, se puede concluir que la decisión de los modelos al clasificar el caso como del grupo de control se debe a la brevedad del audio. En una muestra tan corta, hay menos oportunidades para que la participante manifieste la pérdida de vocabulario, simplificación del lenguaje y otras características que puedan ser registrados por los extractores. Además, a pesar de su baja puntuación en el MMSE, la paciente no presenta fenómenos disfluentes, como pausas o repeticiones de palabras de manera marcada.

Caso 2

En el caso 2 se trata de un hombre que fue clasificado como erróneamente como del grupo de control y cuyo diagnóstico es de Alzheimer.

La transcripción es la siguiente:

> `Well the boy on the chair is falling, reaching up for a cookie, handing one to the girl. The lady is wiping a dish, water running on the floor, she's standing in it. Trees outside, a long shrubbery, a window outside that I can see. That's about it. That's fine.`

Su caso es similar al anterior, en donde se tiene un audio de corta duración al ser de 28 segundos también. Sólo se presentan 2 pausas, una con interjección y otra vacía. En cuanto a la transcripción, el participante cambió la palabra *stool* por una aproximada en el campo semántico *chair* (silla). Se presentan múltiples elipsis del agente y en algunas frases, por ejemplo, *water running on the floor* en vez de *the water is running on the floor*. Pero en general, y a pesar de tener un MMSE de 20, el participante no manifiesta gran cantidad de fenómenos lingüísticos. Por lo tanto, la conclusión es similar a la del caso 1, en la cual la duración corta del audio no da oportunidad a la manifestación de síntomas ni a una correcta caracterización del individuo.

Caso 3

La transcripción del caso 3 es la siguiente:

> Just tell me all of the action. Can you repeat that? Well, the issues with having people around... I don't know what to think of to it yet. You just repeat the sentences after them....that boy on a stool... Mum's mad at them because they broke the window....stool to the elephant......getting cookies... The chalk is on the teacher's desk....mother... The monkey does tricks for us....water running......don't know what they're watching......sink overflowing... Dad called and said he would bring home a pizza for dinner....the curtains are blowing in my......head... We will paint John's wagon green....head......body......head...

Esta transcripción sugiere que se trata de un participante con daño cognitivo serio y que se encuentra en un estado de confusión. Sin embargo, al analizar el audio se detectó que tiene fallas de origen y que presenta el solapamiento de dos voces femeninas. Lo que parece ser interrupciones en la transcripción, en realidad es el intento del modelo de ASR, para transcribir ambas voces de forma intercalada. Por lo tanto, no es posible obtener características de las transcripciones de calidad, lo que hace que el modelo falle. Es importante que la voz en los audios sea clara con el fin de evitar confundir al modelo.

Capítulo 7

Conclusiones y Trabajo Futuro

En este trabajo se presentó una solución basada en la fusión de características de estilo y contenido, con el objetivo de contribuir en la tarea no resuelta de la detección de Alzheimer a partir de voz. Esta tarea consiste en detectar casos de Alzheimer a partir de descripciones de imágenes realizadas por participantes de un estudio médico. Para contribuir con la tarea, se propone un método para abordar el problema desde dos perspectivas distintas, a través de dos grupos de características: uno de contenido, que es de naturaleza semántica; y otra de estilo, que agrupa características léxico-sintácticas y paralingüísticas. Para posteriormente combinarlas, explotando las ventajas que ofrece cada una con el objetivo de mejorar los resultados de clasificación.

Como contribución, se propuso la utilización de características de contenido, una forma alternativa de representar el contenido de las transcripciones, mediante la convolución de *embeddings* con filtros de diferente tamaño. Esto con el fin de describir las transcripciones mediante relaciones de dependencia entre los tókenes, logrando una mejor representación semántica a comparación de la que se obtiene de los embeddings. También, se propuso la utilización de una serie de características de estilo que describen los participantes, mediante estadísticas de categorías gramaticales, complejidad, pausas y disfluencias. A diferencia de otros trabajos, que se limitaron a contar las pausas mediante histogramas, en este trabajo, también se consideraron medidas de dispersión para las pausas y disfluencias. Con el propósito de no solo saber cuantas pausas o disfluencias existen, sino también su distribución a lo largo del diálogo a analizar. Se exploraron formas de combinar características de contenido con estilo, en donde, se consideraron tres estrategias para la combinación, a fin de encontrar aquella que ofrezca la mejor representación. Los métodos explorados fueron los siguientes: una fusión temprana mediante la concatenación simple de características; una fusión utilizando unidades de GMU; y una fusión mediante una selección de modelos basada en centroides.

Se realizaron experimentos por cada modalidad obteniendo los siguientes resultados en términos de exactitud: 0.85 para contenido; 0.86 para estilo. Tanto con las características de contenido como de estilo, se obtienen mejores resultados que con BERT. En especial para las características de contenido, la nueva representación implica una mejor separación entre clases. En el caso de las combinaciones, se obtuvieron las siguientes exactitudes: 0.87 de fusión temprana; 0.86 de fusión con GMU; y 0.87 de fusión mediante selección de clasificadores. En términos generales, todas las fusiones obtuvieron resultados por encima del estado del arte, cuyo mejor resultado fue 0.85, sin mencionar que superó también a todos los *baselines*, por lo tanto, se puede concluir que el mejor método de combinación es la fusión temprana al considerar que no solamente arroja los mejores resultados, sino también su implementación es la más simple.

Se realizaron experimentos y pruebas para analizar con más detalle los resultados. De forma previa a la combinación de resultados se había calculado CFD entre dos clasificadores, uno con características de estilo y otro de contenido, con el propósito de medir el nivel de complementariedad obteniendo un resultado de: 0.83. Un valor de CFD igual a 1 indica que dos clasificadores son distintos, entonces esto implica un grado alto de complementariedad al señalar que los errores en ambos clasificadores no son iguales en su mayoría. Se revisó si esta complementariedad se debía a diferencias en como las características describían, a través de los clasificadores, las instancias de diferentes clases. Y se comprobó que, en efecto, las diferencias se debían a que las características de estilo permiten capturar un 93 % de las instancias de MCI, mientras que las características de contenido un 57 %. Esto sugiere que las características de contenido no permiten una diferenciación adecuada del grupo de control, esto puede deberse a que en los casos de MCI, la parte semántica del habla no se encuentra afectada, a fin de cuentas estos casos se encuentran en la parte superior de la escala del estado cognitivo MMSE, junto a los del grupo de control. Por el otro lado, las características de estilo, que ven únicamente la estructura, logran encontrar diferencias lo suficientemente marcadas como para permitir una mejor detección. Del estudio de ablación de las características de estilo, se observa que son las pausas y difluencias (incluyendo las medidas de dispersión) características relevantes para separar las instancias del grupo de control del resto de clases.

Se buscaron correlaciones entre características de estilo y la escala MMSE de los participantes del estudio. Se obtuvieron algunos resultados acordes a la literatura médica, como la pérdida de sustantivos que contribuye a una puntuación menor. En general, en características con mayor correlación positiva, se encuentran aquellas que reflejan riqueza

en el lenguaje y mejor compresión visual al dar descripciones más detalladas. Entre estas características se encuentran adjetivos, el uso de números cardinales o la diversidad de vocabulario. En cambio, características con una correlación negativa se encuentran todas aquellas que sirven para expresar duda, como los pronombres interrogativos. También, entre las características con mayor correlación negativa se encuentran las medidas de dispersión, es decir, lo cual indica que pausas que se distribuyen a lo largo del diálogo se relacionan con problemas cognitivos.

Para finalizar, el método presentado muestra resultados positivos, aunque no está exento de ciertas vulnerabilidades. El modelo se ve afectado por la calidad de los audios. Factores como el ruido ambiental, voces secundarias y grabaciones de corta duración pueden impactar negativamente la extracción de características. El sonido de fondo puede llevar al modelo a detectar fonemas que no existen; las voces adicionales añaden características que no corresponden al individuo en estudio; y un audio de breve duración no permite que ciertos síntomas se hagan evidentes. Además, el modelo puede confundirse si un paciente, que en términos clínicos es cognitivamente saludable, muestra un habla discontinua debido a una condición no relacionada con Alzheimer o con otras demencias, como podría ser apraxia o un tumor cerebral. También es posible que una persona, a pesar de tener algún deterioro cognitivo, lo experimente en su memoria u otras áreas, pero que en sus habilidades lingüísticas permanezcan sin cambios.

7.1. Trabajo Futuro

Como futuro trabajo, es esencial investigar la aplicabilidad del método propuesto en situaciones donde, en lugar de describir una imagen para obtener audios, se extraigan de entrevistas y diálogos. El contexto presenta varios desafíos adicionales, ya que en conversaciones espontáneas, el tema puede cambiar. Por ende, en esta modalidad, la independencia temática se vuelve crucial, además, es necesario considerar aspectos relacionados con la interacción con el entrevistador, como por ejemplo, pausas largas entre turnos al hablar. Por este motivo, sería necesario modificar el conjunto de características para reflejar tal interacción. Para la parte del conjunto de contenido, es necesario adaptarlo para verificar si existe coherencia entre preguntas y respuestas. En este caso se podría explorar el uso de grandes modelos de lenguaje generativos, por citar un ejemplo, existe LlaMa (Touvron et al. 2023) un modelo que puede ser ejecutado de forma local y soporta *fine-tuning* y entre varios de sus usos, se puede utilizar para verificar si una respuesta contesta una determi-

nada pregunta o si existe duda en una intervención del paciente, etc. De esta manera, se podría explorar de una forma práctica aspectos semánticos que podrían requerir el uso de varios modelos distintos.

Anexos

Tabla 7.1: Categorías Gramaticales.

Categoría	Variantes/Modalidad
Sustantivo	Singular, Plural, Propio
Adjetivo	Base, Comparativo, Superlativo
Adverbio	Base, Comparativo, Superlativo, De partícula
Verbo	Base, Auxiliar Modal, Pasado, Gerundio, Participio Presente, Participio Pasado, Presente 3ra Persona del Singular, Presente no 3ra Persona del Singular
Pronombre	Personal, Posesivo
Signo de Puntuación	

Tabla 7.2: Correlación de Características de Estilo con MMSE.

Característica	Correlación de Pearson	Intensidad
Sustantivo (Singular)	0.4826	Moderada
Conjunción Subordinada	0.3633	Moderada
Determinante	0.3425	Moderada
Prof. media de Árbol Sintáctico	0.3331	Moderada
Adjetivo	0.2851	Débil
Sustantivo (Plural)	0.2325	Débil
Número Cardinal	0.2243	Débil
Pronombre Posesivo	0.1922	Débil
Diversidad de Vocabulario	0.1897	Débil
Verbo (3ra persona, Singular)	0.1881	Débil
Existencial “there”	0.1588	Débil
Verbo (Pasado Participio)	0.1259	Débil
Verbo (Gerundio)	0.0973	Inexistente
Dispersión Global de Pausas	0.0809	Inexistente

Continúa en la siguiente página

Tabla 7.2: Correlación de Características de Estilo con MMSE. (Continuación)

Pausa Mínima	0.0571	Inexistente
Infitivo “To”	0.0458	Inexistente
Signo de Puntuación Superfluo	0.0308	Inexistente
Coma	0.0289	Inexistente
Pausa [500ms, 1s)	-0.0028	Inexistente
Determinante Interrogativo	-0.0069	Inexistente
Marca Posesica	-0.0071	Inexistente
Pausa Promedio	-0.0080	Inexistente
Adjetivo (Superlativo)	-0.0090	Inexistente
Dispersión de Disfluencia [3t, 5t)	-0.0104	Inexistente
Conjunción Coordinada	-0.0147	Inexistente
Pausa STD	-0.0248	Inexistente
Adverbio (Comparativo)	-0.0348	Inexistente
Pausa Máxima	-0.0515	Inexistente
Pausa [5s, ∞)	-0.0679	Inexistente
disfluency_frequency	-0.0749	Inexistente
Adverbio (Interrogativo)	-0.0803	Inexistente
Predeterminante	-0.0975	Inexistente
Guión	-0.1066	Débil
Pausa [2s, 5s)	-0.1073	Débil
Verbo (presente, no 3ra persona)	-0.1099	Débil
Dispersión de Pausa [2s, 5s)	-0.1260	Débil
Adjetivo (Comparativo)	-0.1317	Débil
Verbo (Pasado)	-0.1397	Débil
Cantidad de Pausas	-0.1455	Débil

Continúa en la siguiente página

Tabla 7.2: Correlación de Características de Estilo con MMSE. (Continuación)

Disfluencia [3t, 5t)	-0.1517	Débil
Verbo (Auxiliar)	-0.1528	Débil
Pausa Total	-0.1580	Débil
Dispersión Global de Disfluencias	-0.1608	Débil
Sustantivo (Singular, Propio)	-0.1609	Débil
Dispersión de Pausa [5s, ∞)	-0.1645	Débil
Dispersión de Disfluencia [1, 3)	-0.1811	Débil
Dispersión de Pausa [500s, 1s)	-0.1842	Débil
Pausa [1s, 2s)	-0.1951	Débil
Verbo (Base)	-0.2216	Débil
Disfluencia [5t, ∞)	-0.2271	Débil
Signo de Puntuación (... / . / ?)	-0.2324	Débil
Disfluencia [1t, 3t)	-0.2487	Débil
Puntos Suspensivos	-0.2775	Débil
Dispersión de Pausa [1s, 2s)	-0.2860	Débil
Adverbio de Partícula	-0.3248	Moderada
Interjección	-0.3597	Moderada
Pronombre Personal	-0.3739	Moderada
Adverbio	-0.3744	Moderada
Pronombre Interrogativo Personal	-0.4881	Moderada

Tabla 7.3: Ganancia de Información.

Característica	Ganancia de Información
Pausa Total	0.9983
Dispersión de Pausa [500ms, 1s)	0.9983

Continúa en la siguiente página

Tabla 7.3: Ganancia de Información. (Continuación)

Pausa Mínima	0.9983
Cantidad de Pausas	0.9863
Diversidad de Vocabulario	0.9622
Pausa [1s, 2s)	0.9513
Pronombre Personal	0.9410
Prof. media de Árbol Sintáctico	0.9260
Signo de Puntuación (... / . / ?)	0.9185
Sustantivo (Singular)	0.9163
Determinante	0.9094
Verbo (3ra persona, Singular)	0.9019
Dispersión de Pausa [1s, 2s)	0.8937
Conjunción Subordinada	0.8762
Adverbio	0.8687
Pause [2s, 5s)	0.8348
Adverbio de Partícula	0.8310
Adjetivo	0.7913
Interjección	0.7335
Pronombre Posesivo	0.6600
Dispersión de Pausa [2s, 5s)	0.6305
Pronombre Interrogativo Personal	0.5465
Existencial “there”	0.5239
Número Cardinal	0.4473
Disfluencia [1t, 3t)	0.4340
Dispersión Global de Pausas	0.2053
Dispersión de Pausa [5s, ∞)	0.1738
Puntos Suspensivos	0.1702

Continúa en la siguiente página

Tabla 7.3: Ganancia de Información. (Continuación)

Dispersión de Disfluencia [1t, 3t)	0.1635
Disfluencia [5t, ∞)	0.1193
Disfluencia [2, 5)	0.1008

Bibliografía

Juarez-Cedillo, Teresa et al. (ene. de 2022). "Prevalence of Dementia and Main Subtypes in Mexico: The Study on Aging and Dementia in Mexico (SADEM)". En: *Journal of Alzheimer's Disease* 89.3, págs. 931-941. ISSN: 1387-2877. DOI: 10.3233/JAD-220012.

Vigo, Inês, Luis Coelho y Sara Reis (ene. de 2022a). "Speech- and Language-Based Classification of Alzheimer's Disease: A Systematic Review". En: *Bioengineering* 9.1, pág. 27. ISSN: 2306-5354. DOI: 10.3390/bioengineering9010027. URL: https://www.mdpi.com/2306-5354/9/1/27.

Yang, Qin et al. (dic. de 2022). "Deep Learning-based Speech Analysis for Alzheimer's Disease Detection: A Literature Review". En: *Alzheimer's Research and Therapy* 14.1, págs. 1-16. ISSN: 17589193. DOI: 10.1186/S13195-022-01131-3/FIGURES/4. URL: https://alzres.biomedcentral.com/articles/10.1186/s13195-022-01131-3%20http://creativecommons.org/publicdomain/zero/1.0/.

Luz, Saturnino, Fasih Haider, Sofia de la Fuente et al. (2020). "Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge". En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2020-October, págs. 2172-2176. ISSN: 19909772. DOI: 10.21437/INTERSPEECH.2020-2571.

Luz, Saturnino, Fasih Haider, Sofia De La Fuente et al. (2021). "Detecting Cognitive Decline Using Speech-only: The ADReSSochallenge". En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 6, págs. 4211-4215. ISSN: 19909772. DOI: 10.21437/INTERSPEECH.2021-1220.

Bird, Helen et al. (jun. de 2000). "The Rise and Fall of Frequency and Imageability: Noun and Verb Production in Semantic Dementia". En: *Brain and Language* 73.1, págs. 17-49. ISSN: 0093-934X. DOI: 10.1006/BRLN.2000.2293.

Martinc, Matej y Senja Pollak (2020). "Tackling the ADReSS Challenge: A Multimodal Approach to the Automated Recognition of Alzheimer's Dementia". En: DOI: 10.21437/

Interspeech.2020-2202. URL: http://dx.doi.org/10.21437/Interspeech.2020-2202.

Lyon, Pamela et al. (mar. de 2021). "Reframing Cognition: Getting Down to Biological Basics". En: *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 376.1820. ISSN: 1471-2970. DOI: 10.1098/RSTB.2019.0750. URL: https://pubmed.ncbi.nlm.nih.gov/33487107/.

Graff-Radford, Jonathan y Angela M. Lunde (2020). *Mayo Clinic on Alzheimer's disease and other dementias*. Ed. por Paula M. Marlow Limbeck. Mayo Clinic Press, pág. 414. ISBN: 9781893005617, 9780795352928.

Folstein, Marshal F., Susan E. Folstein y Paul R. McHugh (1975). "Mini-mental state. A Practical Method for Grading the Cognitive State of Patients for the Clinician". En: *Journal of Psychiatric Research* 12.3, págs. 189-198. ISSN: 00223956. DOI: 10.1016/0022-3956(75)90026-6.

Gallo, Jennifer L. (ago. de 2013). "Contemporary Intellectual Assessment: Theories, Tests, and Issues". En: *Archives of Clinical Neuropsychology* 28.5, págs. 507-508. ISSN: 0887-6177. DOI: 10.1093/ARCLIN/ACT011. URL: https://dx.doi.org/10.1093/arclin/act011.

Hutt, Max L. (abr. de 2007). "Bender-Gestalt Test." En: *The Genain quadruplets: A case study and theoretical analysis of heredity and environment in schizophrenia.*, págs. 241-256. DOI: 10.1037/11420-015.

Zaudig, Michael (1992). "A New Systematic Method of Measurement and Diagnosis of "Mild Cognitive Impairment" and Dementia According to ICD-10 and DSM-III-R Criteria". En: *International Psychogeriatrics* 4.4, págs. 203-219. ISSN: 1741-203X. DOI: 10.1017/S1041610292001273. URL: https://www.cambridge.org/core/journals/international-psychogeriatrics/article/abs/new-systematic-method-of-measurement-and-diagnosis-of-mild-cognitive-impairment-and-dementia-according-to-icd10-and-dsmiiir-criteria/27DD028B57B6FC3FCFA4C74358CC7178.

Global action plan on the public health response to dementia 2017-2025 (2017). Inf. téc. Geneva: World Health Organization.

Yuan, Jiahong et al. (2020). "Disfluencies and Fine-tuning Pre-trained Language Models for Detection of Alzheimer's Disease". En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2020-October, págs. 2162-2166. ISSN: 19909772. DOI: 10.21437/INTERSPEECH.2020-2516.

López-de-Ipiña, Karmele et al. (mayo de 2013). "On the Selection of Non-Invasive Methods Based on Speech Analysis Oriented to Automatic Alzheimer Disease Diagnosis". En: *Sensors 2013, Vol. 13, Pages 6730-6745* 13.5, págs. 6730-6745. ISSN: 1424-8220. DOI: 10.3390/S130506730. URL: https://www.mdpi.com/1424-8220/13/5/6730/htm%20https://www.mdpi.com/1424-8220/13/5/6730.

Lou, Paria Jamshid y Mark Johnson (2020). "Improving Disfluency Detection by Self-Training a Self-Attentive Model". En: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, págs. 3754-3763. ISSN: 0736587X. DOI: 10.18653/V1/2020.ACL-MAIN.346. URL: https://aclanthology.org/2020.acl-main.346.

Loring, David W. y Stephen Bowden (2015). *INS dictionary of neuropsychology and clinical neurosciences*. 2da, pág. 397. ISBN: 9780195366457. URL: https://global.oup.com/academic/product/ins-dictionary-of-neuropsychology-and-clinical-neurosciences-9780195366457.

Wu, Qi et al. (2022). "Visual Question Answering". En: Advances in Computer Vision and Pattern Recognition. DOI: 10.1007/978-981-19-0964-1. URL: https://link.springer.com/10.1007/978-981-19-0964-1.

Zafar, Iffat. et al. (2018). "Hands-On Convolutional Neural Networks with TensorFlow : Solve Computer Vision Problems with Modeling in TensorFlow and Python." En: pág. 264.

Severyn, Aliaksei y Alessandro Moschittiy (ago. de 2015). "Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks". En: *SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, págs. 373-382. DOI: 10.1145/2766462.2767738. URL: https://dl.acm.org/doi/10.1145/2766462.2767738.

Miner, Gary et al. (ene. de 2012). "Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications". En: *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*, págs. 1-1053. DOI: 10.1016/C2010-0-66188-8. URL: http://www.sciencedirect.com:5070/book/9780123869791/practical-text-mining-and-statistical-analysis-for-non-structured-text-data-applications.

Ravichandiran, Sudharsan (2021). *Getting Started with Google BERT: Build and train state-of-the-art natural language processing models using BERT*. Packt Publishing Ltd.

Mikolov, Tomas et al. (ene. de 2013). "Efficient Estimation of Word Representations in Vector Space". En: *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. URL: https://arxiv.org/abs/1301.3781v3.

Wang, Lipo (2004). "Soft Computing in Communications". En: Studies in Fuzziness and Soft Computing 136. DOI: 10.1007/978-3-540-45090-0. URL: http://link.springer.com/10.1007/978-3-540-45090-0.

Li, Jinyu et al. (2015). *Robust Automatic Speech Recognition*. Elsevier. ISBN: 9780128023983. DOI: 10.1016/C2014-0-02251-4.

Sen, Soumya, Anjan Dutta y Nilanjan Dey (2019). *Audio Processing and Speech Recognition*. Singapore: Springer Singapore. ISBN: 978-981-13-6097-8. DOI: 10.1007/978-981-13-6098-5.

Radford, Alec et al. (2023). "Robust Speech Recognition via Large-Scale Weak Supervision". En: *ICML'23: Proceedings of the 40th International Conference on Machine Learning*. URL: https://github.com/openai/.

Vaswani, Ashish et al. (2017). "Attention Is All You Need". En: *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*. DOI: 10.5555/3295222.3295349. URL: https://dl.acm.org/doi/10.5555/3295222.3295349.

Balagopalan, Aparna, Benjamin Eyre et al. (2020). "To BERT or not to BERT: Comparing Speech and Language-based Approaches for Alzheimer's Disease Detection". En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2020-October, págs. 2167-2171. ISSN: 19909772. DOI: 10.21437/INTERSPEECH.2020-2557.

Pye, Clifton y Brian MacWhinney (mar. de 1994). "The CHILDES Project: Tools for Analyzing Talk". En: *Language* 70.1, pág. 156. ISSN: 00978507. DOI: 10.2307/416745.

Tóth, László et al. (2015). "Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech using ASR". En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2015-January, págs. 2694-2698. ISSN: 19909772. DOI: 10.21437/INTERSPEECH.2015-568.

Pan, Yilin, Bahman Mirheidari, Markus Reuber et al. (2020). "Improving Detection of Alzheimer's Disease using Automatic Speech Recognition to Identify High-quality Segments for more Robust Feature Extraction". En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2020-October, págs. 4961-4965. ISSN: 19909772. DOI: 10.21437/INTERSPEECH.2020-2698.

Becker, James T. et al. (jun. de 1994). "The Natural History of Alzheimer's Disease: Description of Study Cohort and Accuracy of Diagnosis". En: *Archives of Neurology* 51.6, págs. 585-594. ISSN: 0003-9942. DOI: 10.1001/ARCHNEUR.1994.00540180063015. URL: https://jamanetwork.com/journals/jamaneurology/fullarticle/592905.

Mirheidari, Bahman et al. (mayo de 2019). "Computational Cognitive Assessment: Investigating the Use of an Intelligent Virtual Agent for the Detection of Early Signs of Dementia". En: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 2019-May, págs. 2732-2736. ISSN: 15206149. DOI: 10.1109/ICASSP.2019.8682423.

Warnita, Tifani, Nakamasa Inoue y Koichi Shinoda (mar. de 2018). "Detecting Alzheimer's Disease Using Gated Convolutional Neural Network from Audio Data". En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2018-September, págs. 1706-1710. ISSN: 19909772. DOI: 10.21437/Interspeech.2018-1713. URL: https://arxiv.org/abs/1803.11344v1.

Pan, Yilin, Bahman Mirheidari, Markus Reuber et al. (2019). "Automatic Hierarchical Attention Neural Network for Detecting AD". En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2019-September, págs. 4105-4109. ISSN: 19909772. DOI: 10.21437/INTERSPEECH.2019-1799.

Balagopalan, Aparna, Ksenia Shkaruta y Jekaterina Novikova (nov. de 2020). "Impact of ASR on Alzheimer's Disease Detection: All Errors are Equal, but Deletions are More Equal than Others". En: págs. 159-164. DOI: 10.18653/V1/2020.WNUT-1.21. URL: https://aclanthology.org/2020.wnut-1.21.

Papineni, Kishore et al. (2002). "BLEU: A Method for Automatic Evaluation of Machine Translation". En: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, págs. 311-318. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040.

Rohanian, Morteza, Julian Hough y Matthew Purver (2020). "Multi-modal Fusion with Gating Using Audio, Lexical and Disfluency Features for Alzheimer's Dementia Recognition from Spontaneous Speech". En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2020-October, págs. 2187-2191. ISSN: 19909772. DOI: 10.21437/INTERSPEECH.2020-2721.

Zhu, Youxiang et al. (mayo de 2021). "Exploring Deep Transfer Learning Techniques for Alzheimer's Dementia Detection". En: *Frontiers in Computer Science* 3, pág. 22. ISSN: 26249898. DOI: 10.3389/FCOMP.2021.624683/BIBTEX.

Schneider, Steffen et al. (2019). "Wav2vec: Unsupervised Pre-training for Speech Recognition". En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2019-September, págs. 3465-3469. ISSN: 19909772. DOI: 10.21437/INTERSPEECH.2019-1873.

Zhang, Zhengyan et al. (2019). "ERNIE: Enhanced Language Representation with Informative Entities". En: *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, págs. 1441-1451. DOI: 10.18653/V1/P19-1139. URL: https://aclanthology.org/P19-1139.

Sarawgi, Utkarsh et al. (2020). "Multimodal Inductive Transfer Learning for Detection of Alzheimer's Dementia and its Severity". En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2020-October, págs. 2212-2216. ISSN: 19909772. DOI: 10.21437/INTERSPEECH.2020-3137.

Pan, Yilin, Bahman Mirheidari, Jennifer M. Harris et al. (2021). "Using the Outputs of Different Automatic Speech Recognition Paradigms for Acoustic- and BERT-based Alzheimer's Dementia Detection through Spontaneous Speech". En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 6, págs. 4216-4220. ISSN: 19909772. DOI: 10.21437/INTERSPEECH.2021-1519.

Panayotov, Vassil et al. (ago. de 2015). "Librispeech: An ASR Corpus based on Public Domain Audio Books". En: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 2015-August, págs. 5206-5210. ISSN: 15206149. DOI: 10.1109/ICASSP.2015.7178964.

Carletta, Jean et al. (2006). "The AMI Meeting Corpus: A Pre-announcement". En: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 3869 LNCS, págs. 28-39. ISSN: 03029743. DOI: 10.1007/11677482{_}3. URL: https://dl.acm.org/doi/10.1007/11677482_3.

Gratch, Jonathan et al. (2014). *The Distress Analysis Interview Corpus of human and computer interviews*. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf.

Manohar, Vimal, Daniel Povey y Sanjeev Khudanpur (ene. de 2018). "JHU Kaldi System for Arabic MGB-3 ASR Challenge Using Diarization, Audio-transcript Alignment and Transfer Learning". En: *2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017 - Proceedings* 2018-January, págs. 346-352. DOI: 10.1109/ASRU.2017.8268956.

Novikova, Jekaterina (2021). "Robustness and Sensitivity of BERT Models Predicting Alzheimer's Disease from Text". En: *W-NUT 2021 - 7th Workshop on Noisy User-Generated Text, Proceedings of the Conference*, págs. 334-339. DOI: 10.18653/V1/2021.WNUT-1.37. URL: https://aclanthology.org/2021.wnut-1.37.

Nasreen, Shamila, Julian Hough y Matthew Purver (2021). "Detecting Alzheimer's Disease Using Interactional and Acoustic Features from Spontaneous Speech". En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 1, págs. 306-310. ISSN: 19909772. DOI: 10.21437/INTERSPEECH.2021-1526.

Levinson, Stephen C. (jun. de 1983). *Pragmatics*. Cambridge University Press. ISBN: 9780511813313. DOI: 10.1017/CBO9780511813313. URL: https://www.cambridge.org/highereducation/books/pragmatics/6D0011901AE9E92CBC1F5F21D7C598C3#contents.

Eyben, Florian, Martin Wöllmer y Björn Schuller (2010). "OpenSMILE - The Munich Versatile and Fast Open-source Audio Feature Extractor". En: *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference*, págs. 1459-1462. DOI: 10.1145/1873951.1874246. URL: https://dl.acm.org/doi/10.1145/1873951.1874246.

Syed, Muhammad Shehram Shah et al. (2020). "Automated Screening for Alzheimer's Dementia through Spontaneous Speech". En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2020-October, págs. 2222-2226. ISSN: 19909772. DOI: 10.21437/INTERSPEECH.2020-3158.

Kim, Yoon (2014). "Convolutional Neural Networks for Sentence Classification". En: *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, págs. 1746-1751. DOI: 10.3115/V1/D14-1181. URL: https://aclanthology.org/D14-1181.

Vigo, Inês, Luis Coelho y Sara Reis (ene. de 2022b). "Speech- and Language-Based Classification of Alzheimer's Disease: A Systematic Review". En: *Bioengineering 2022, Vol. 9, Page 27* 9.1, pág. 27. ISSN: 2306-5354. DOI: 10.3390/BIOENGINEERING9010027. URL: https://www.mdpi.com/2306-5354/9/1/27/htm%20https://www.mdpi.com/2306-5354/9/1/27.

Arevalo, John et al. (jul. de 2020). "Gated Multimodal Networks". En: *Neural Computing and Applications* 32.14, págs. 10209-10228. ISSN: 14333058. DOI: 10.1007/S00521-019-04559-1. URL: https://dl.acm.org/doi/10.1007/s00521-019-04559-1.

Bergstra, James et al. (2011). "Algorithms for Hyper-Parameter Optimization". En: *Advances in Neural Information Processing Systems* 24.

Pérez-Toro, P. A. et al. (2022). "Alzheimer's Detection from English to Spanish Using Acoustic and Linguistic Embeddings". En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2022-September, págs. 2483-2487. ISSN: 19909772. DOI: 10.21437/INTERSPEECH.2022-10883.

Baevski, Alexei et al. (2020). "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". En: *Advances in Neural Information Processing Systems* 33, págs. 12449-12460. URL: https://github.com/pytorch/fairseq.

WebRTC: Real-Time Communication in Browsers (s.f.). URL: https://www.w3.org/TR/webrtc/.

Matthews, B. W. (oct. de 1975). "Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme". En: *Biochimica et Biophysica Acta (BBA) - Protein Structure* 405.2, págs. 442-451. ISSN: 0005-2795. DOI: 10.1016/0005-2795(75)90109-9.

Powers, D M W y Ailab (oct. de 2020). "Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation". En: URL: https://arxiv.org/abs/2010.16061v1.

Wang, Wenjia (2008). "Some Fundamental Issues in Ensemble Methods". En: *Proceedings of the International Joint Conference on Neural Networks*, págs. 2243-2250. DOI: 10.1109/IJCNN.2008.4634108.

Syed, Zafi Sherhan et al. (2021). "Tackling the ADRESSO Challenge 2021: The MUET-RMIT System for Alzheimer's Dementia Recognition from Spontaneous Speech". En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 6, págs. 4231-4235. ISSN: 19909772. DOI: 10.21437/INTERSPEECH.2021-1572.

Rohanian, Morteza, Julian Hough y Matthew Purver (2021). "Alzheimer's Dementia Recognition using Acoustic, Lexical, Disfluency and Speech Pause Features Robust to Noisy Inputs". En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 6, págs. 4191-4195. ISSN: 19909772. DOI: 10.21437/INTERSPEECH.2021-1633.

Pappagari, Raghavendra et al. (2021). "Automatic Detection and Assessment of Alzheimer Disease using Speech and Language Technologies in Low-resource Scenarios". En: *Proceedings of the Annual Conference of the International Speech Communication Associa-*

tion, INTERSPEECH 6, págs. 4206-4210. ISSN: 19909772. DOI: 10.21437/INTERSPEECH.2021-1850.

Chen, Jun et al. (2021). "Automatic Detection of Alzheimer's Disease Using Spontaneous Speech-only". En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 6, págs. 4181-4185. ISSN: 19909772. DOI: 10.21437/INTERSPEECH.2021-2002.

Touvron, Hugo et al. (feb. de 2023). "LLaMA: Open and Efficient Foundation Language Models". En: URL: https://arxiv.org/abs/2302.13971v1.