UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA

INSTITUTO PARA EL DESARROLLO TECNOLÓGICO Y LA INNOVACIÓN EN COMUNICACIONES (IDETIC)

PROGRAMA DE DOCTORADO

EMPRESA, INTERNET Y TECNOLOGÍAS DE LAS COMUNICACIONES

TESIS DOCTORAL

Avances en el análisis del habla mediante sistemas conversacionales automáticos aplicados a la enfermedad de Alzheimer

Dña. María Luisa Barragán Pulido

Las Palmas de Gran Canaria
Diciembre 2021

UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA

INSTITUTO PARA EL DESARROLLO TECNOLÓGICO Y LA INNOVACIÓN EN COMUNICACIONES (IDETIC)

PROGRAMA DE DOCTORADO

EMPRESA, INTERNET Y TECNOLOGÍAS DE LAS COMUNICACIONES

TESIS DOCTORAL

Avances en el análisis del habla mediante sistemas conversacionales automáticos aplicados a la enfermedad de Alzheimer

AUTORA: Dña. María Luisa Barragán Pulido

DIRECTORES: Dr. D. Jesús Bernardino Alonso Hernández

Dr. D. Miguel Ángel Ferrer Ballester

El Director **El Codirector** **La Doctoranda**

Las Palmas de Gran Canaria a 8 de diciembre de 2021.

A mis padres y abuelos

Agradecimientos

En primer lugar, quisiera agradecer a mis directores de tesis, Dr. Jesús Bernardino Alonso Hernández y Dr. Miguel Ángel Ferrer Ballester, por haberme ayudado a hacer posible este trabajo.

Al Dr. Jesús Bernardino Alonso Hernández, a quien admiro profundamente, quiero agradecerle, en especial, por su paciencia y también por haberme enseñado, aunque no pueda reflejarse en estas páginas, una verdadera actitud de vida.

Por último, quisiera agradecer a mis padres, a mis hermanas y a mi novio, Osinori, por haberme acompañado incondicionalmente en este camino.

Resumen

La enfermedad de Alzheimer es actualmente, como la han catalogado algunos expertos, la epidemia del presente siglo. En una sociedad cada vez más envejecida y en la que se espera las cifras de afectados se tripliquen en 2050, se presenta casi como una urgencia disponer de un mayor conocimiento e investigar en torno a una enfermedad de la que todavía se desconocen sus causas y, como todos sabemos, para la que no existe cura conocida.

Entre los múltiples biomarcadores actualmente bajo investigación, el habla se presenta como un potente indicador del estado cognitivo y emocional de una persona. En base a esta evidencia, esta tesis plantea como hipótesis que es posible discriminar la enfermedad de Alzheimer a partir de la voz obtenida mediante métodos conversacionales automáticos. A partir de aquí y en línea con las tendencias actuales en la eSalud y la Telemedicina se propone el uso de este tipo de sistemas teniendo en cuenta las numerosas ventajas que presenta respecto a sus homólogos clásicos basados en entrevistadores humanos, implícitamente más subjetivos y menos escalables.

Para conocer hasta qué punto el habla obtenida a partir de métodos conversacionales automáticos tiene potencial en el desarrollo de herramientas no invasivas, rápidas, objetivas, escalables y de bajo coste que, llegado el caso, faciliten o contribuyan a la detección precoz de la enfermedad de Alzheimer, se ha empleado la base de datos CSAP-R19, novedosa con respecto a las localizadas en el campo debido a que está formada por dos tipos de muestras para cada sujeto entrevistado: grabaciones recogidas con los métodos tradicionales (entrevistador humano) y muestras obtenidas mediante el *software* Prognosis (entrevistador automático).

A partir de estas grabaciones de voz se ha identificado y extraído una serie de características temporales y medidas de carga emocional basadas, estas últimas, en la Temperatura Emocional. Posteriormente se han llevado a cabo diferentes análisis estadísticos univariantes y multivariantes, así como estudios basados en técnicas de Aprendizaje Máquina (*Machine Learning*) enfocados siempre a validar la hipótesis planteada.

Los resultados que hemos obtenido son prometedores y, en lo que respecta al potencial de las herramientas conversacionales automáticas aplicadas a la detección precoz de la enfermedad de Alzheimer, abren nuevas vías de investigación en torno al desarrollo y mejora de este tipo de métodos. Asimismo, se espera contribuyan a dar un impulso dentro de una realidad que, a día de hoy, ya es la eSalud o la Telemedicina.

Tabla de contenidos

Índice de figuras

Índice de tablas

Glosario

AD	*Alzheimer's Disease*, enfermedad de Alzheimer
AD1	Enfermedad de Alzheimer grado leve
AD2	Enfermedad de Alzheimer grado moderado
AB	*AdaBoost M1*
AE	*Approximate Entropy,* entropía aproximada
ALS	*Amyotrophic Lateral Sclerosis*, Esclerosis Lateral Amiotrópica (ELA)
ANN	*Artificial Neural Network*, red neuronal artificial
APOE	*Apolipoprotein*, Apolipoproteína
ASSA	*Automatic Spontaneous Speech Analysis*, análisis automático de habla espontánea
BN	*Bayesian Network*, red bayesiana
CD	*Correlation Dimension*, dimensión de correlación
CE	*Correlation Entropy*, entropía de correlación
CLAN	*Computer Language Analysis*, análisis de lenguaje computacional
CNN	*Convolutional Neural Network*, red neuronal convolucional
CSF	*Cerebrospinal Fluid*, fluido cerebroespinal
CT	*Computerized Tomography*, tomografía computarizada
cVF	*Verbal Fluency categories*, categorías de fluidez verbal
DAT	*Dementia of Alzheimer Type*, demencia de tipo Alzheimer
DFA	*Detrended Fluctuation Analysis*, análisis de fluctuación sin tendencia
DL	*Deep Learning*, Aprendizaje Profundo
EEG	*Electroencephalography*, electroencefalografía
ERA	*Emotional Response Analysis,* análisis de la respuesta emocional
ESA	*Emotional Speech Analysis*, análisis del habla emocional
F0	*Fundamental Frequency*, frecuencia fundamental o *pitch*
FD	*Fractal Dimension*, dimension fractal
FMMI	*First Minimum of Mutual Information*, primer mínimo de información mutua
GCNN	*Gated Convolutional Neural Network,* red neuronal convolucional cerrada
HC	*Healthy Control*, sujeto de control
HE	*Hurst Exponent*, Exponente de Hurst
HNR	*Harmonic to Noise Rate*, relación armónico a ruido
ICU	*Information Content Unit*, unidad de contenido de información
kNN	*k Nearest Neighbours*, k vecinos más cercanos
LDA	*Linear Discriminant Analysis*, análisis discriminante lineal
LLE	*Largest Lyapunov Exponent*, mayor exponente de Lyapunov
LPC	*Linear Prediction Coefficient*, coeficiente de predicción lineal
MANOVA	*Multivariate analysis of variance*, análisis multivariante de la variaza
MB	*Meta Bagging*
MCI	*Mild Cognitive Impairment*, desorden cognitivo leve

MEG	*Magnetoencefalography*, magnetoencefalografía
MFCC	*Mel Frequency Cepstral Component*, componente cepstral Mel-Frequency
MFDFA	*Multifractal Disactivated Fluctuation Analysis*, Análisis de fluctuación multifractal desactivado
ML	*Machine Learning*, Aprendizaje Máquina
MLP	*Multilayer Perceptron*, perceptron multicapa
MMSE	*Mini Mental State Examination*, Mini examen del estado mental
MRI	*Magnetic Resonance Image*, imagen de resonancia magnética
MSE	*Mean Square Error*, error cuadrático medio
NB	*Naive Bayes*
NHR	*Noise to Harmonic Rate*, relación ruido a armónico
NLP	*Natural Language Production*, producción del lenguaje natural
NN	*Neural Network*, red neuronal
PD	*Parkinson's Disease*, enfermedad de Parkinson
PET	*Positron Emission Tomography*, tomografía de emisión de positrones
PID	*Propositional Idea Density*, densidad de ideas proposicionales
PPA	*Primary Progressive Aphasia*, afasia progresiva primaria
RBE1	*First order Rényi Block Entropy*, entropía de bloque Rényi de primer orden
RBE2	*Second order Rényi Block Entropy*, entropía de bloque Rényi de segundo orden
RE	*Second order Rényi Entropy*, entropía de Rényi de segundo orden
RF	*Random Forest*
RNN	*Recurrent Neural Network*, red neuronal recurrente
RPDE	*Recurrence Probability Density Entropy*, entropía de densidad de probabilidad de recurrencia
RST	*Rethoric Structure Theory*, teoría de la estructura retórica
SD	*Standard Deviation*, desviación estándar
SE	*Sample Entropy*, Entropía de muestra
SEO	*Square Energy Operator*, operador de energía cuadrado
SHE	*Shannon Entropy*, entropía de Shannon
SMO	*Sequential Minimal Optimization*, optimización minima secuencial
SPECT	*Simple Photon Emission Computed Tomography*, tomografía computarizada por emisión de fotones simple
SS	*Spontaneous Speech*, habla espontánea
SVM	*Support Vector Machine*, máquina de vectores soporte
TE	Temperatura Emocional
TEO	*Teager Kaiser Energy Operator*, operador de energía de Teager Kaiser
VAD	*Voice Activity Detector*, detector de actividad de voz
VR	*Variation Range*, rango de variación
ZL	*Ziv Lempel Complexity*, complejidad de Ziv Lempel

Capítulo 1 Introducción

1.1 Antecedentes

La enfermedad de Alzheimer (en adelante EA o AD, por sus siglas en inglés) es, actualmente, la causa más común de demencia neurodegenerativa en el mundo. Supone entre el 70-76% de los casos de demencia en los países desarrollados, cada vez más longevos [1], [2]. Aunque la etiología de la AD es desconocida y es probable que se trate de una enfermedad de causa multifactorial, se sabe que su inicio es insidioso y que aparece en la edad adulta, produciendo fundamentalmente el deterioro cognitivo y conductual de la persona [1]. El daño que produce, es progresivo e irreversible; no existe cura y conlleva, a todos los efectos, deterioro y muerte neuronal [3]. La pérdida de memoria aparece como uno de los primeros síntomas, al que se suman también dificultades con el uso del lenguaje y la capacidad para realizar actividades diarias o, incluso, en estados más avanzados, dificultades para llevar a cabo funciones corporales básicas como caminar o tragar [4]. Finalmente, el paciente pierde por completo su autonomía y requiere la atención constante de un cuidador o familiar. Este hecho sumado al lento desarrollo de la enfermedad que suele ser de unos 8-10 años desde el inicio hasta la muerte [4], [5], la cual se produce normalmente por alguna otra enfermedad intercurrente, hace que estos cuidados se prolonguen en el tiempo.

Algunos estudios consideran la AD como una de las enfermedades de mayores consecuencias sociales y económicas tanto por la carga social y que a la salud pública supone, como al daño personal y familiar que produce en el paciente [3], [5]. Otros muchos son los que la consideran la epidemia del presente siglo [1]; en 2050 el número de personas con demencia, síndrome clínico de la AD, triplicará las cifras de 2010. Cifras que se doblan cada 20 años [6].

Entre las causas del incremento de prevalencia de AD, destacan la inversión de la pirámide generacional y el aumento de la longevidad [5], directamente relacionado con los avances en la asistencia médica, así como a las condiciones sociales y ambientales. Debido a que uno de los principales factores de riesgo es la edad, la

prevalencia es más alta en mujeres que en hombres [7], [8], las cuales tienen una mayor esperanza de vida. A partir de los 65 años, en cualquier caso, el riesgo de padecer Alzheimer se duplica cada cinco años [9].

Por otra parte, debido a la asistencia médica y los avances, así como a las condiciones sociales y ambientales el número de personas mayores de entre 80-90 años o más se espera que crezca dramáticamente en los países occidentales.

Por su parte, la AD se ha convertido en una de las enfermedades crónicas más costosas de la sociedad [10]. En comparación con otras enfermedades, los gastos producidos por personas con AD y otras demencias fueron más de dos veces el de personas de la misma edad con cáncer y un 74% mayor respecto a personas con enfermedades cardiovasculares [11], [12].

En 2015, sólo los costes directos de atención médica y social a nivel mundial alcanzaron los $487 mil millones, lo que supone un 0,65 % del total de los productos interiores brutos (PIB) - un enorme impacto económico para un sólo grupo de trastornos, especialmente dado que el 87% de los costes se producen en los países de ingresos más altos. Según datos del Registro Sueco de Demencia (SveDem), sólo los costes de diagnóstico de un caso de AD en 2010 en Suecia rondaron los 5.500 € [13].

A parte de los gastos iniciales de diagnóstico y farmacia, se suma el coste de los cuidados de personas con demencia, siendo ésta la parte más importante. Según un estudio realizado en 2016 los cuidados sociales (asistencia en el hogar a largo plazo) supusieron el 42,3% de los gastos y los informales (proporcionados por cuidadores no profesionales), el 41,7%. Ambos cuidados representan la mayor proporción de costes frente a la atención médica directa, del 16% [13].

Es importante destacar que los costes dependen directamente del grado de severidad de la enfermedad (grado leve, moderado y severo) así como del país de residencia y sexo del paciente, por ejemplo, existe un mayor gasto en farmacia para mujeres con grado moderado de AD [14]. Por su parte se ha estimado que en 2015 el mercado mundial de productos farmacéuticos y de diagnóstico para la AD suponía aproximadamente sólo el 1% de los costes totales de la enfermedad, hecho que por otro lado pondría de manifiesto la ausencia de una terapia efectiva [13].

Como ejemplo, en Reino Unido pacientes en estados más avanzados atendidos en casa podrían alcanzar los 26.800 € de gasto al año frente a los 44.100 y 46.260 €/año de Suiza y España, respectivamente [15]. En el caso concreto de países como España los pacientes y/o las familias asumen más del 60% del coste global [14], lo cual supone un gasto medio de 30.500 €/año. Teniendo en cuenta que en España la pensión media a día 1 de junio de 2017 es de 11.042,64 €/año (según datos del Ministerio de Empleo y Seguridad Social del Gobierno de España), el coste medio de esta asistencia triplicaría las pensiones recibidas.

A día de hoy, cuidar a una persona con AD u otra demencia plantea difíciles desafíos, especialmente cuando la personalidad y el comportamiento del paciente se ven afectados [16]. Es a partir de ahí, a medida que la enfermedad progresa y los síntomas empeoran, cuando se requieren niveles crecientes de supervisión y cuidado, generando sobre los familiares un mayor estrés emocional y depresión; problemas de salud nuevos o exacerbados [5]. Además de esto, se añade un terrible daño sobre la intimidad, las

experiencias compartidas y recuerdos que a menudo son parte de la relación cuidador-paciente y que se ven amenazadas debido al deterioro cognitivo y funcional que acompañará, inexorablemente, a la AD [5]. Según un estudio estadounidense, los cuidadores de personas con demencia dedican de media 27 horas/semana más que los cuidadores de personas sin demencia y, el 26% de ellos, 41 horas más [17], [18].

Dado el impacto de esta enfermedad, incrementar las ayudas económicas existentes, potenciar las ayudas domésticas, la atención médica y los servicios de enfermería a domicilio, crear centros para estancias diurnas, incluso incentivar a la familia con deducciones fiscales, podrían ser algunas de las soluciones a corto plazo a ofrecer a las familias [19].

De mano de la innovación tecnológica también se plantean interesantes soluciones enfocadas al apoyo a personas con demencia y sus cuidadores. La tendencia apunta a mejorar los aspectos de la seguridad de las personas con demencia por medio de dispositivos como los detectores de humo o botones de pánico, la mejora de determinadas capacidades como la memoria, dispositivos de sistemas de posicionamiento global o mensajes de voz, principalmente para personas en las primeras etapas de la demencia [20]. Otras herramientas se centran en el tratamiento y los aspectos terapéuticos para personas con demencia y sus cuidadores, por ejemplo mediante videoteléfonos y asesoramiento en línea [3], [21], [22] o por medio de videojuegos cuya finalidad es la ejercitación física (*exergames*) [20], [23]. Algunos de ellos son ComputerLink, AlzOnline o Caring-for-Others [24].

El uso de dispositivos como tabletas u otras herramientas y aplicaciones como terapia en la AD parece ser un recurso eficaz con múltiples implicaciones prácticas [25]. Según diversos estudios que han realizado una revisión sistemática de la eficacia de este tipo de herramientas, determinadas intervenciones tuvieron efectos moderados para mejorar el estrés y la depresión del cuidador y, además, observaron que los resultados variaban con sus características personales (grupos étnicos, apoyo formal recibido o carga basal). En cualquier caso, se requieren más estudios para definir qué intervenciones son más efectivas en situaciones específicas [24], [23]. Actualmente, el desarrollo de herramientas mejoradas para "personalizar" los servicios y que los cuidadores maximicen sus beneficios representan un área de investigación emergente [24], [26]–[28].

Por su parte, el uso de la tecnología podría ser una solución no sólo a nivel de apoyo a pacientes y cuidadores sino también en la búsqueda de métodos no-invasivos, económicos, rápidos y fácilmente aplicables [29] para la detección precoz y control evolutivo de la enfermedad. El número de publicaciones realizadas en los últimos años en torno al *eHealth* y al *Telecare* se ha incrementado notablemente. En la Figura 1-1 se relaciona el número de publicaciones realizadas al respecto, distinguiendo aquéllas publicadas en revistas Q1.

Actualmente, se han llevado a cabo algunas iniciativas internacionales a gran escala como el Programa Conjunto de la UE para la Investigación de Enfermedades Neurodegenerativas (JPND), la cumbre sobre la demencia del G8 [30], la Iniciativa de Medicamentos Innovadores o la Organización para la Cooperación y el Desarrollo Económico para la obtención de *Big Data* en la investigación sobre AD [31]. Por su

parte, varios países en Europa ya han iniciado grandes ensayos de intervención multifactorial basados en el estilo de vida. Algunos de estos ensayos completados o en curso son FINGER [32], MAPT [33], PreDIVA [34] y HATICE [35].

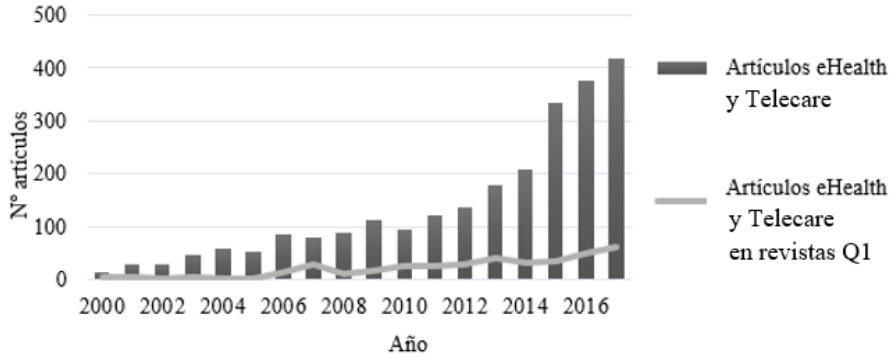


Figura 1-1 Publicaciones Q1 centradas en *eHealth* y *Telecare* y realizadas desde el año 2000 al 2017 [36].

Lo que, en cualquier caso, parece claro es que sólo los aumentos dirigidos a la inversión en investigación proporcionarán alguna esperanza de encontrar una cura para la AD, desarrollar estrategias para retrasar la aparición o ralentizar la progresión de la AD e identificar medidas preventivas y terapéuticas eficaces que puedan aplicarse en diferentes contextos [13].

1.1.1 La enfermedad de Alzheimer

La AD abarca el proceso completo de los cambios patológicos iniciales en el cerebro antes de que los síntomas aparezcan a través de la demencia [3]. Esto significa que la AD incluye no sólo aquellos pacientes con demencia debido a la enfermedad, sino también aquellos con deterioro cognitivo leve (en adelante, MCI) y a personas asintomáticas que han verificado tener biomarcadores positivos de la AD. Bajo las directrices actuales, la demencia pasa a considerarse un síndrome causado por la acumulación de cambios cerebrales producidos por la AD [3], [5].

Los cambios cerebrales comienzan a producirse porque las células nerviosas (neuronas) en determinadas partes del cerebro involucradas en la función cognitiva han sido dañadas o destruidas. El ritmo al que los síntomas avanzan de leve a moderado y, posteriormente, a severo, varía con la persona [5] y cursa con mayor velocidad cuando es diagnosticado a edades tempranas.

El síntoma temprano más común de la demencia es la dificultad para recordar acontecimientos recientes. A medida que el trastorno se desarrolla, surgen otros síntomas como desorientación, cambios de humor, incluyendo apatía, depresión, ansiedad, agitación, alteraciones del sueño y confusión. A medida que la AD avanza las pérdidas de memoria y los cambios de comportamiento se hacen más notables, aparecen dificultades para hablar y comunicarse y disminución de la capacidad motriz. Llega un momento en el curso de la enfermedad en el que el paciente tiene problemas para desarrollar las funciones más básicas de la vida diaria como comer y tragar [13]. La acumulación progresiva de la discapacidad, con deterioro en múltiples dominios cognitivos, interfiere sustancialmente con el funcionamiento diario, incluyendo el desarrollo social y familiar del paciente [5], [13].

Inicialmente, los cambios cognitivos que produce son muy sutiles, poco claros y difíciles de detectar y diferenciar de otras patologías [5]. Con el avance de la enfermedad y a medida que estos síntomas se agravan, aparecen otros tipos de enfermedades intercurrentes que, por norma general, producen la muerte del paciente. La neumonía es la principal causa de muerte en pacientes de AD. En otros casos las defunciones pueden producirse incluso por desnutrición y deshidratación [5].

Aunque actualmente no se conocen las causas exactas por las que una persona desarrolla AD sí se han podido relacionar algunos de los factores de riesgo y de protección que podrían afectar [37], si bien su relevancia está abierta al debate. Los factores de riesgo más pronunciados son el avance de la edad y el transporte de uno o dos alelos APOE ε4. También influye [13] el riesgo vascular y factores metabólicos, el estilo de vida, la nutrición y la depresión.

Hasta hoy, el diagnóstico de la AD [38] en pacientes vivos se ha basado casi exclusivamente en exámenes clínicos, con confirmación neuropatológica post mortem, la cual se realiza sólo en ocasiones y normalmente en el contexto de la investigación [39]. Actualmente, los médicos se limitan a diagnosticar AD probable después de descartar otras causas potenciales de deterioro cognitivo y confirmando deterioros en funciones o actividades de la vida diaria [40].

Atendiendo a la terminología clásica [41]–[43], se definen diferentes estados; el deterioro cognitivo preclínico, leve (MCI) y la demencia de AD, si bien, el Grupo de Trabajo International para Nuevos Criterios de Investigación de AD [44] ha propuesto un nuevo léxico de diagnóstico para describir la progresión de la AD.

En el pasado, el principal síntoma de la enfermedad, la demencia de AD, era difícil de diferenciar de otras patologías. Gracias el advenimiento de técnicas avanzadas de neuroimagen, como las imágenes de resonancia magnética (MRI) o la tomografía de emisión de positrones (PET), se ha alcanzado una comprensión más amplia en los procesos neuropatológicos de los pacientes. Además de pruebas de neuroimagen y exámenes médicos, se incluyen diferentes test neuropsicológicos, análisis de líquido cefalorraquídeo (CSF) y análisis de sangre [29].

El proceso de diagnóstico sigue siendo complejo y se realiza inevitablemente en las fases avanzadas de la enfermedad [1], se alarga en el tiempo y supone costes elevados. Requiere, además, de médicos especialistas para poder llevarlo a cabo. Trabajos anteriores realizados en EEUU [45] reflejan una preocupante falta de facultativos y detectan en los médicos no-especialistas limitaciones para identificar de manera precisa AD temprana y MCI [29], [45], ambos, hitos significativos de la AD.

Actualmente, dentro de los diferentes tipos de diagnóstico clínico, los análisis de líquido CSF y las pruebas de neuroimagen son los principales y más relevantes. Mediante el análisis de líquido cefalorraquídeo es posible determinar proteínas implicadas en la AD (beta-amiloide 42, tau total y tau fosforilada) y, a través de MRI y PET, realizar mediciones del hipocampo y del córtex entorrinal (tanto glucosa como amiloide) [1]. El hecho de que se trate de métodos caros e invasivos, limita su uso como herramienta de screening [29] y mantiene despierto el interés en la búsqueda de biomarcadores en lugares más accesibles del cuerpo, que pudieran ser sensibles a la AD antes de la aparición clínica de la demencia [47].

El uso de biomarcadores accesibles como método de screening se presenta, por tanto, como una solución económica para el diagnóstico precoz de la AD preclínica [46], y la identificación y supervisión de las etapas específicas de la AD [47]. Actualmente, este campo constituye un área activa de investigación [47] poniendo especial énfasis en el desarrollo de algunos como es el sanguíneo. Desafortunadamente, ningún biomarcador plasmático ha tenido hasta ahora la especificidad y sensibilidad suficiente [47].

No son pocas las investigaciones que apuntan a test clínicos basados en biomarcadores obtenidos a través de la evaluación subjetiva de la memoria, la depresión tardía, el análisis de habla, del olfato o de la manera de caminar. De manera relativamente reciente, se encuentran bajo estudio los test neurofisiológicos basados en electroencefalografías (EEG) y magnetoencefalografías (MEG) [29]. Hasta la fecha no se han obtenido resultados contundentes.

Más allá de estas investigaciones centradas en la búsqueda de biomarcadores y de su utilidad como método de *screening*, la realidad actual es que los pacientes todavía deben presentar deterioro cognitivo apreciable o déficits funcionales concomitantes para ser diagnosticados y, en consecuencia, considerar la intervención farmacológica [40]. Cuando esto ocurre, lo más probable es que ya se hayan producido daños en el tejido cerebral y pérdidas irreversibles [48], [49]. Actualmente se comercializan medicamentos que pueden mejorar el funcionamiento de la memoria [50], aunque éstos no alteran la enfermedad [50]–[52].

La mayoría de los ensayos clínicos de nuevos fármacos se ha basado en el estudio de pacientes con AD leve a moderada, y parece confirmarse que las terapias neuroprotectoras aplicadas durante las primeras etapas de la enfermedad conducirían a mejores resultados clínicos en los individuos [53]. Actualmente, los fármacos empleados para el tratamiento de la AD mejoran temporalmente los síntomas aumentando la cantidad de neurotransmisores en el cerebro, aunque su efectividad varía con la persona.

Muchos factores dificultan el desarrollo de tratamientos más eficaces. Entre ellos, el alto coste de desarrollo y el tiempo relativamente largo necesario para observar si un tratamiento en investigación afecta a la progresión de la enfermedad y la estructura del cerebro, la cual está protegida por la barrera hematoencefálica, a través de la cual, además, sólo pueden emplearse medicamentos especializados [3]. En el caso de las terapias no farmacológicas, no se ha podido demostrar todavía que alteren el curso de la AD, aunque sí han encontrado que algunos métodos como el ejercicio aeróbico y la estimulación cognitiva pueden resultar beneficiosos [54].

Mientras que los avances en el campo del tratamiento se quedan por detrás de otros subcampos de la investigación de la AD, parece claro que, una vez se obtenga un mayor conocimiento sobre la patología de la AD y su relación molecular con otras partes del cerebro, se realizará un progreso adicional en el frente del tratamiento [47].

1.1.2 El habla en la enfermedad de Alzheimer

Los problemas de lenguaje son considerados uno de los síntomas más característicos de la enfermedad de Alzheimer, los cuales aparecen como consecuencia directa e inevitable del deterioro cognitivo [55]. Existe evidencia de que las personas con MCI con impedimentos en varios dominios, además del comunicativo, son más propensos a desarrollar AD [56], [57]. La afasia progresiva primaria (APP), síndrome clínico caracterizado por un deterioro progresivo del lenguaje de etiología neurodegenerativa, podría ser el modo de presentación de diferentes enfermedades neurodegenerativas como la AD [58].

Tal y como se ha demostrado, los problemas de lenguaje y comunicación son evidentes durante las interacciones entre pacientes y neurólogos, y las observaciones interaccionales podrían usarse para diferenciar entre dificultades cognitivas debidas a trastornos neurodegenerativos o trastornos de memoria funcional [59]. Los estudios que han demostrado el desempeño deficiente que presentan los pacientes de AD en diferentes pruebas lingüísticas [60] son numerosos y algunos han encontrado una relación significativa entre las puntuaciones de test como el *Minimental State Examination* (MMSE) y diferentes medidas de lenguaje realizadas (articulación, fluidez semántica, repetición y denominación (anomia)) [61]. Asimismo, se ha podido constatar que las personas con AD presentan más dificultades en la denominación, repiten ideas con más frecuencia, hacen uso de pausas vacías más prolongadas, usan un lenguaje más simplificado, presentan una prosodia monótona y un discurso menos informativo, coherente y cohesionado que otros grupos de control [62]. En la Tabla 1-1 se relacionan estas alteraciones en los diferentes campos lingüísticos y su evolución en las fases individuales de la AD [61].

Tabla 1-1 Efectos de la enfermedad de Alzheimer sobre los aspectos lingüísticos del habla de pacientes AD [61].

Campos afectados		**MCI**	**AD leve**	**AD moderada**	**AD severa**	**Conclusiones parciales**
Fonético-fonológico		++	++	+++	++++	Cambios temporales en el SS. Incremento en el número y tiempo de vacilación Parafasia fonémica.
Procesamiento sintáctico		+	+	++	++++	Reduce la complejidad sintáctica en AD moderado Introduce errores gramaticales en AD severa.
Procesamiento léxico-semántico		++	++	+++	++++	Factor predictivo positivo de demencia. Indicador fiable del deterioro del lenguaje. Distingue AD de HC. Individuos con MCI normalmente presentan déficits.
Fluidez	**Semántica**	++	++	+++	++++	Deficiencias en MCI (mayor en verbos que en sustantivos).
	Fonológica	++	++	+++	++++	No muy común en MCI y AD temprano.
Carga del discurso		+	++	+++	++++	Textos más cortos, menos información relevante y múltiples tipos de error.

MCI: *Mild Cognitive Impairment,* AD: *Alzheimer's Disease,* HC: *Healthy Control,* SS: *Spontaneous Speech, símbolo (+): indicador del grado de afección.*

Los problemas comunicativos específicos, como la afasia o la anomia, dependen de la etapa de la enfermedad [29] y aumentan con el curso de la AD [61], [63], [64]. Años antes del establecimiento del diagnóstico clínico, el lenguaje ya muestra un deterioro cognitivo significativo en pacientes preclínicos [65], [66]. Concretamente, algunas fuentes exponen que en el primer año después del comienzo de la enfermedad los diferentes aspectos del lenguaje aparecen obscurecidos por la pérdida de interés y espontaneidad, desorientación espacial y trastornos de la memoria [67]–[69]. Aunque esto afecta a la fluidez verbal, a menudo no se detecta. La capacidad de respuesta emocional se ve afectada y se observan cambios sociales y de comportamiento [41] que podrían deberse a esta pérdida de memoria. Asimismo, la alteración de las habilidades de percepción podrían magnificar algunas respuestas emocionales [29]. En el intervalo que va del primer al tercer año del comienzo de la AD, los déficit del lenguaje ya se hacen más notables [67], [68]; se requiere mayor tiempo para encontrar las palabras en el discurso y se presenta mayor cantidad de frases vacías e indefinidas, las cuales son más cortas, menos complejas y con menor variedad de entonación. También, se corrigen en menor medida los errores [70]. Cuando la enfermedad está más avanzada, en las fases moderada y severa, se produce un rápido decremento de fluidez verbal dando lugar, finalmente, a la descomposición de la comprensión [61].

En base a lo anterior, algunos estudios afirman que la AD puede ser detectada más sensiblemente con la ayuda de un análisis lingüístico que con otros exámenes cognitivos [61]. Sin embargo, la pregunta que se plantea es qué características del lenguaje y del habla podrían identificar la enfermedad en su etapa inicial [61].

Por otra parte, las emociones son características fundamentales de los seres humanos, afectan a la percepción y las actividades cotidianas, como la comunicación y la toma de decisiones. Se expresan a través del habla, expresiones faciales, gestos y otras pistas no verbales. La diferencia de estados emocionales puede considerarse también como uno de los criterios de evaluación importantes para medir el rendimiento del procedimiento de cognición [71], [72].

1.1.3 Exámenes lingüísticos tradicionales

Con objeto de identificar y cuantificar el grado de deterioro cognitivo de un paciente de AD se han utilizado múltiples pruebas lingüísticas [73]. Algunas de las capacidades evaluadas han sido la identificación, comprensión, repetición o lectura. Estas capacidades se alteran de manera evidente en los estadios leve a severo, si bien pueden localizarse ligeras variaciones específicas en el caso de alguna de ellas [73]. Por su parte el habla espontánea ha sido ampliamente utilizada mediante tareas de descripción de imágenes, vídeos o conversaciones. Según demuestran algunos estudios [74], a partir de tareas de descripción de habla es posible distinguir entre AD y controles así como identificar otros aspectos cognitivos.

Se han definido test estandarizados como el Token Test [75] (evalúa el procesamiento semántico), Test de Boston [76] (también evalúa el procesamiento semántico y se basa en pruebas de nominación) y el Test de Fluidez Verbal/Nominación (la más utilizada es la prueba de Fluidez Verbal por Categorías (VFc) que explora la

agrupación y cambio de categorías). También existen baterías que incluyen, cada uno en su modalidad, varias pruebas lingüísticas (nominación, comprensión, repetición o lectura) y no lingüísticas (como orientación espacial y temporal, atención y cálculo, memoria, recuerdo inmediato, dibujo y escritura) [77]. Algunos de ellos son el *Minimental State Examination* de Folstein (MMSE) [77], el Test de Cuetos-Vega o el Evaluación Cognitiva de Montreal, aunque no son los únicos.

El MMSE, a pesar de ser el más extendido y utilizado en la práctica clínica, tiene ciertas limitaciones, como baja sensibilidad a las primeras etapas de la AD, incapacidad para diferenciar el tipo de demencia, baja especificidad y sesgo cultural por nivel educativo [77].

Las líneas más tradicionales y extendidas se han centrado durante años en estos exámenes de tipo neuropsicológico, identificando y evaluando manualmente los síntomas cognitivos, conductuales y comunicativos más sobresalientes y tempranos [61], [78], [79]. El hecho de que se trate de pruebas de sencilla aplicación, rapidez y sensibilidad, ha provocado que se desarrollen numerosos test y se haya perdido cualquier tipo de consenso [2] a la hora de establecer exámenes clínicos generales. Actualmente, la tendencia en cualquiera de las áreas bajo investigación exige definir métodos automatizables y objetivos [80].

1.2 Retos

Actualmente los métodos de diagnóstico aplicados a la enfermedad de Alzheimer son caros, invasivos y difíciles de extender a toda la población. En este complicado contexto, el habla aparece como un potente indicador del estado cognitivo de pacientes de Alzheimer, resultando afectada incluso años antes del diagnóstico clínico. Por este motivo, disponer de técnicas como puede ser el análisis automático del habla (por ejemplo, como método de *screenning* o para documentar objetivamente el estado de un paciente) resulta de especial interés.

Yendo más allá del mero análisis automático del habla, una herramienta conversacional automatizada, además de ser aplicable a gran escala e independiente de los entrevistadores que a día de hoy conducen las entrevistas, presenta múltiples ventajas como solución eSalud (*eHealth*, por su terminología en inglés). Entre otras, su escalabilidad, su bajo coste y su aplicación en entornos remotos.

1.3 Hipótesis

En esta tesis se plantea como premisa principal que, a partir de un sistema conversacional automatizado, es posible discriminar pacientes AD frente a sujetos sanos obteniendo resultados similares a los de otros métodos de grabación y entrevistas tradicionales. Asimismo, identificamos las siguientes premisas adicionales:

- Un sistema conversacional automatizado es una herramienta escalable, económica, no invasiva y no dependiente de una persona que guía la entrevista.

Dota, entre otras cosas, de anonimato al sujeto entrevistado y resulta ser un método cómodo capaz de realizarse en el entorno del paciente, en cualquier momento y horario. Un entrevistador automático debe, además, reunir condiciones que humanicen la herramienta, desarrollando tanto la empatía con el sujeto como la utilización de recursos o estímulos que permitan acercarse al entrevistado.

- A partir de las grabaciones de habla obtenidas de un sistema conversacional automatizado pueden extraerse diferentes características cuantificables de la voz sensibles a la enfermedad de Alzheimer (por ejemplo, cambios en los tiempos de habla y silencio, cambios en los aspectos frecuenciales o, también, léxicos y semánticos, entre otros).
- Las características temporales y de Temperatura Emocional, no siendo las únicas, aportan, mediante técnicas de procesado de voz, información relevante sobre el estado cognitivo de un sujeto y suficiente a la hora de discriminar AD.
- Dentro del concepto de la eSalud o *eHealth*, la automatización del proceso de entrevistas enfocada a la detección de la enfermedad de Alzheimer ofrece soluciones automatizadas, más objetivas, escalables, no invasivas y de bajo coste, en comparación con los métodos de entrevistas tradicionales.

1.4 Objetivos

El objetivo general de esta tesis es conocer si a partir de un sistema conversacional automático es posible discriminar pacientes AD de locutores sanos a partir de la voz y, además, obtener resultados similares a los obtenidos históricamente mediante los métodos de entrevistas tradicionales. En este sentido, se plantea evaluar de forma conjunta y novedosa las diferentes metodologías de grabación de muestras, así como diversos aspectos temporales y emocionales de la voz, ofreciendo un enfoque del no constan referencias en trabajos similares. Los objetivos específicos son:

- **O1. Herramienta *software***
  Contribuir al desarrollo de una herramienta *software* que permita obtener grabaciones de manera automática y que posteriormente pasarán a formar parte de nuestra base de datos. Analizar y definir los diferentes aspectos que debe reunir esta herramienta automática y las condiciones necesarias para conseguir la correcta grabación de dichas muestras.

- **O2. Base de datos**
  Crear una base de datos con muestras obtenidas mediante un sistema automático de grabación, así como con grabaciones obtenidas manualmente, por el método tradicional. Además, para cada tipo de método de grabación, recoger tanto muestras de voces sanas como patológicas AD.

- **O3. Extracción de características del habla**
  A partir de las muestras de la base de datos creada, obtener mediante el uso de técnicas de procesado automático de voz, diferentes características del habla que han demostrado en otros estudios ser sensibles a la AD. Concretamente, se plantea extraer un conjunto de características temporales y otras medidas relacionadas con la carga emocional, tanto para voces sanas como para patológicas y tanto para el entrevistador automático como para el humano.

- **O4. Análisis del habla**
  Analizar y cuantificar aspectos de la voz como las duraciones de habla y silencio o la carga emocional basándonos en diferentes análisis estadísticos y técnicas de Aprendizaje Máquina (*Machine Learning*). A partir de estos análisis, identificar patrones de similitud entre las grabaciones obtenidas automáticamente y las obtenidas manualmente. Sumado a esto, identificar dichos patrones y fenómenos también entre los diferentes tipos de sujetos analizados o poblaciones, concretamente: sujetos sanos y pacientes con diferentes grados de severidad de la enfermedad.

1.5 Metodología

La metodología seguida en esta tesis se ha estructurado en cuatro grandes bloques. El primero de ellos, centrado en las tareas de documentación y revisión crítica del estado del arte, desde información concerniente a aspectos fundamentales de la AD (como la situación e impacto social, económico y sanitario) hasta información centrada concretamente en el procesado automático de voz aplicado a la detección precoz y/o control evolutivo de la AD, pasando por los actuales métodos de diagnóstico o los diferentes análisis lingüísticos llevados a cabo en medicina hasta la fecha, entre otros.

Realizada esta revisión se ha contribuido a desarrollar y mejorar el software de grabación automática de muestras. Se han analizado y definido diferentes aspectos necesarios para su correcto funcionamiento y, posteriormente, las muestras obtenidas a partir de él se han utilizado para componer nuestra base de datos.

A partir de este análisis, el siguiente paso ha sido la contribución e incorporación de estas muestras automáticas a la base de datos empleada en esta tesis Cross-Sectional Alzheimer Prognosis R2019 (en adelante, CSAP-R19) realizando tareas de adecuación, tratamiento y análisis de datos. Asimismo, a fin de poder realizar nuestro estudio comparativo, también se han incorporado a la base de datos muestras obtenidas manualmente.

Con la base de datos ya conformada, se han llevado a cabo las tareas de extracción de características, basándonos en aspectos temporales y emocionales de la voz que previamente han demostrado ser sensibles a la presencia de la AD. Posteriormente, se han llevado a cabo diferentes análisis estadísticos y entrenamiento y prueba de modelos de clasificación de Aprendizaje Máquina, con el objetivo de buscar

diferencias y similitudes en cuanto a la capacidad discriminativa (AD) de cada método de grabación analizado: automático vs manual.

A continuación en la Figura 1-2 se presenta esquemáticamente la metodología empleada para las diferentes muestras, entrevistadores y características analizadas.

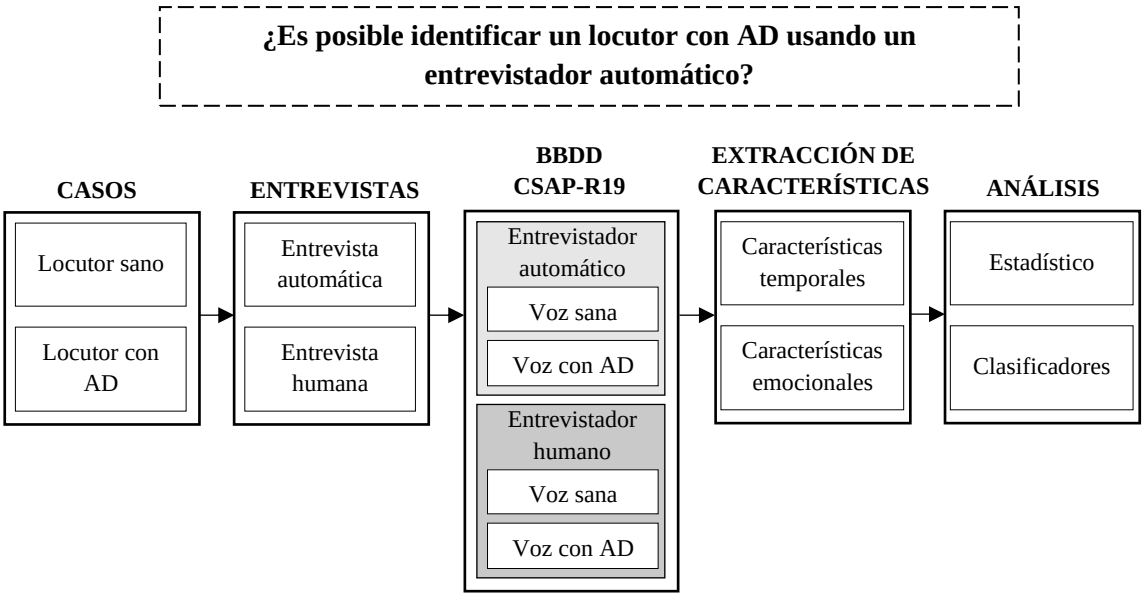


Figura 1-2 Metodología empleada para la conformación de la base de datos CSAP-R19, la extracción de características y análisis de las muestras.

1.6 Contribuciones

Como resultado de esta tesis, se han publicado varias contribuciones de investigación en revistas científicas JCR:

- **Artículo 1.** Publicado a partir de las revisiones realizadas en el Capítulo 1 y Capítulo 2 de esta memoria. En esta publicación se presenta un estudio crítico sobre el estado de la técnica en lo que respecta al procesado automático de voz aplicado a la detección precoz y/o control evolutivo de la AD.

  M. L. Barragán-Pulido, J. B. Alonso-Hernández, Ferrer-Ballester; M. A., C. M. Travieso-González, J. Mekyska, and Z. Smékal, "Alzheimer's disease and automatic speech analysis: a review," Expert Syst. Appl., vol. 150, p. 113213, 2020.

  Citas recibidas: 31 (Fuente Google scholar: consulta realizada el 08/12/21).

  Expert Systems with Applications (ISSN 0957-4174) con índice de impacto ISI-JCR 6,954 (en 2020):

  - *Computer Science, Artificial Intelligence*: Q1 (23/149).
  - *Engineering, Electrical & Electronic*: Q1 (24/273). Nota: primer decil.
  - *Operations Research & Management Science*: Q1 (8/84). Nota: primer decil.

- **Artículo 2.** Publicado a partir del Capítulo 4, Capítulo 5 y Capítulo 7 de esta memoria. Se relacionan y presentan de manera crítica las bases de datos localizadas en el campo, sus principales características y su disponibilidad. Asimismo, se presenta el sistema conversacional automático *software* Prognosis y la base de datos desarrollada a partir de él, CSAP-R19, especialmente novedosa desde el punto de vista del tipo de muestras que la conforman: grabaciones de sujetos sanos y pacientes de AD con diferentes grados de severidad, así como muestras de habla obtenidas de un sistema conversacional automático y también por los métodos tradicionales. Posteriormente, basándonos en un proceso de extracción de características temporales y análisis estadísticos descriptivos y no paramétricos, como primera aproximación, se publican unos resultados que reflejan de manera preliminar la capacidad discriminativa del entrevistador automático. A partir de ellos se compara su rendimiento en comparación con su homólogo humano.

  Jesús B. Alonso-Hernández, María L. Barragán-Pulido, Jose M. Gil-Bordón, Miguel Ángel Miguel Ángel Ferrer-Ballester, Carlos M. Travieso-Gonzalez, Using a human interviewer or an automatic interviewer in the evaluation of patients with AD from speech?, Applied Sciences, Ed. MDPI AG, (doi: 10.3390/app11073228), (ISSN: 2076-3417), vol 11, num. 2021. Índice de impacto ISI-JCR: 2,679 (en 2020) Q2 (32/91).

  Citas recibidas: 1 (Fuente Google scholar: consulta realizada el 08/12/21).

  Applied Sciences (ISSN 2076-3417) con índice de impacto ISI-JCR 2,679 (en 2020):
  - *Engineering, Multidisciplinary:* Q2 (38/90).
  - *Physics, Applied:* Q2 (73/160).
  - *Chemistry, Multidisciplinary:* Q3 (84/219).
  - *Materials Science, Multidisciplinary:* Q3 (201/334).

- **Artículo 3.** Redactado a partir de los resultados obtenidos en el Capítulo 7 y Capítulo 8 de esta memoria. A partir de la base de datos CSAP-R19 se lleva a cabo un proceso de extracción de características temporales que posteriormente son analizadas mediante diferentes estudios estadísticos univariantes, paramétricos y no paramétricos, y multivariantes. Se analizan las características más relevantes y se aplican diferentes análisis discriminantes basados en cuatro modelos de clasificación. Actualmente este artículo se encuentra aceptado por la revista con índice de impacto Expert Systems with Applications (ESWA) y está prevista su inmediata publicación en las próximas semanas.

  Jesús B. Alonso-Hernández, María L. Barragán-Pulido, Jose M. Gil-Bordón, Miguel Ángel Miguel Ángel Ferrer-Ballester, Carlos M. Travieso-Gonzalez, Speech Evaluation of patients with Alzheimer's Disease using an automatic interviewer.

  Expert Systems with Applications (ISSN 0957-4174) con índice de impacto ISI-JCR 6,954 (en 2020).

- **Artículo editorial.** Redactado a partir de los resultados del Capítulo 2, se expone de manera resumida la situación actual del procesado automático del habla aplicado a la enfermedad de Alzheimer, así como los nuevos retos a los que nos enfrentamos a día de hoy en el campo.

  J. B. Alonso Hernández and M. L. Barragán Pulido, "Alzheimer's Disease: New Challenges for Speech Analysis," Int. J. Comput. Softw. Eng., vol. 4, no. 2, Jun. 2019.

1.7 Estructura de la memoria

La memoria comienza con este primer capítulo, en el que se exponen los antecedentes, así como los retos e hipótesis que llevan a realizar esta investigación, objetivos, contribuciones principales y un breve resumen de la estructura de la memoria.

En el Capítulo 2 se presenta la revisión crítica del estado del arte centrada en el análisis automático del habla y la enfermedad de Alzheimer. Para contextualizar la situación actual de la enfermedad y la importancia de este tipo de análisis, se comienza haciendo un repaso los rasgos principales que esta patología presenta, haciendo un recorrido sobre los síntomas más evidentes y, posteriormente, se analiza el impacto social y económico de la misma. Se presentan asimismo las diferentes afecciones que la enfermedad de Alzheimer provoca en el lenguaje, haciendo de este uno de sus síntomas más característicos. También se describen los diferentes tipos de exámenes lingüísticos empleados en la actualidad como métodos de diagnóstico para la detección de la enfermedad. Por último, centrados en los diferentes análisis automáticos basados en habla y detección de AD, se realiza una clasificación de las principales características lingüísticas empleadas, así como una revisión de todos los trabajos identificados que han ido empleando este tipo de técnicas automáticas desde que se tiene constancia.

A continuación, se presenta Capítulo 3 donde se describe detalladamente la metodología seguida en la tesis. Éste se divide en diferentes apartados, cada uno correspondiente a un capítulo de la memoria.

El Capítulo 4 tiene como objetivo presentar la base de datos CSAP-R19 compuesta por diferentes tipos de muestras tomadas de manera automática (entrevistador automático) y manual (entrevistador humano). También se hace un repaso a las principales bases de datos relacionadas con la AD que hemos podido localizar, detallando cada una de ellas.

En los dos siguientes capítulos, Capítulo 5 y Capítulo 6, se describe el proceso de extracción de características temporales y de Temperatura Emocional llevado a cabo sobre la base de datos CSAP-R19, definiendo y formulando cada una de las medidas extraídas.

En el Capítulo 7 se presentan los resultados obtenidos de los diferentes análisis estadísticos realizados sobre las características temporales y de Temperatura Emocional descritas en los capítulos anteriores y obtenidas de las muestras que conforman la base de datos CSAP-R19. Se presentan los resultados de diferentes estudios univariantes y multivariantes basados en técnicas de estadística descriptiva, estudios paramétricos y no paramétricos que nos permiten cuantificar las diferencias existentes entre el habla de

pacientes de Alzheimer y sujetos de control, así como disponer de información relativa a la influencia que el método de grabación (entrevistador humano y automático) puede tener en el proceso discriminatorio.

En el Capítulo 8 se presentan los resultados obtenidos de aplicar sobre las muestras varios modelos de clasificación con el fin de conocer si el entrevistador automático iguala o no la capacidad discriminativa del entrevistador humano a partir de las características temporales y de Temperatura Emocional. En este capítulo, a diferencia del anterior, se realizan exclusivamente análisis multivariantes con las diferentes medidas extraídas y varios modelos de clasificación, además de un proceso previo de selección de características.

El Capítulo 9 incluye el análisis y discusión sobre el estado del arte y bases de datos relacionadas, así como una interpretación crítica del proceso de extracción de características y el posterior análisis estadístico. Se presenta asimismo una discusión en torno a los resultados obtenidos del proceso de clasificación llevado a cabo, selección de características, y entrenamiento y prueba de modelos.

La memoria finaliza con el Capítulo 10 donde se exponen las principales conclusiones extraídas y una propuesta de diferentes líneas futuras y mejoras a esta tesis.

Capítulo 2 Revisión del estado del arte: la enfermedad de Alzheimer y el procesado automático de voz

En este capítulo se ha llevado a cabo una revisión crítica del estado del arte en la que, primeramente, se ha tratado de contextualizar de manera amplia y de poner en valor las numerosas investigaciones que se están llevando a cabo en torno a esta enfermedad. De la misma manera, se han descrito los principales aspectos comunicativos de la AD, poniendo el foco fundamentalmente sobre el procesado automático del habla aplicado a la detección de Alzheimer.

Para ello, se ha realizado una revisión sistemática del estado del arte a través de herramientas de búsqueda bibliográfica como SpringerLink, Scopus, ScienceDirect y Google Academic. Los criterios utilizados para localizar los trabajos relacionados se han basado en la búsqueda de ciertos conceptos de interés, yendo desde los más amplios hasta los más específicos. Concretamente se buscaron las palabras "demencia" y "Alzheimer" en un primer paso; y, a partir de aquí, se delimitaron los resultados filtrando sucesivamente por "Alzheimer y procesamiento/procesado", "Alzheimer y procesamiento/procesado y automático" y, finalmente, por "Alzheimer y procesamiento/procesado y automático y voz o habla". Estos conceptos, tanto en el título como en el resumen de los documentos debían, al menos, aparecer. Todos los resultados localizados y presentados a continuación han sido examinados manualmente.

2.1 Procesado de voz en la enfermedad de Alzheimer

Desde hace unos años, las técnicas que se basan en el procesado automático de la señal de voz a partir de su grabación han encontrado un importante nicho en la evaluación del lenguaje aplicado a la detección precoz de enfermedades neurodegenerativas. Estas técnicas ofrecen la posibilidad de cuantificar propiedades de la señal que resultan

relevantes en la descripción de una determinada patología. Posteriormente, y apoyadas en técnicas de Aprendizaje Automático, se realiza el proceso de clasificación de las muestras en función de las características o medidas obtenidas. Estos métodos presentan la ventaja de poder aplicarse de manera automática evitando la posible influencia del intermediario o entrevistador.

Las características empleadas más convencionales en detección de AD han sido las lineales, clínicamente interpretables. Por su parte, otras corrientes han hecho uso de técnicas no lineales, más recientes y novedosas. Ambas suponen un importante indicador de las estructuras expresivas del habla y son dependientes de las tareas realizadas [81]. Además de los procesos de extracción de características, donde se combinan análisis estadísticos y/o modelados matemáticos [81], para cuantificar las propiedades del habla, se añade el uso de algoritmos de Aprendizaje Máquina (*Machine Learning*) en el proceso de clasificación.

Los modelos de clasificación, a partir de resultados concretos y objetivos que no son el fin último, aportan *explicabilidad* al fenómeno real que subyace detrás y que buscamos entender. Nos permiten realizar una interpretación, en este caso, sobre la afección que la AD tiene sobre la voz, añadiendo un valor extra al uso de estas técnicas. Los métodos de clasificación automática se dividen en dos fases fundamentales: entrenamiento y validación. Hasta ahora se han diseñado diferentes tipos de clasificadores para llevar a cabo esta tarea. Algunos de ellos son *Support Vector Machines* (en adelante, SVM), *k-Nearest Neighbors* (en adelante, kNN) o *Neural Networks* (en adelante, NN), cada uno con sus características de diseño y funcionamiento específicas. En este mismo contexto, el concepto de *Deep Learning* aparece como un método de aprendizaje automático más complejo que los anteriores y que también se está abriendo paso en el ámbito de la detección automática de Alzheimer a partir de la voz.

El uso de métodos de clasificación que utilizan características lingüísticas resultantes de emisiones verbales podría facilitar el control evolutivo de pacientes con AD para una amplia población. En los últimos años, se ha documentado un incremento casi exponencial del número de investigaciones al respecto con el objeto de introducir el análisis del habla como un biomarcador no invasivo de AD [81], si bien por ahora no han sido incluidos en la práctica clínica.

Por su parte, la tendencia actual muestra avances en empresas importantes como IBM Watson [82], Cantab [83], Winterlights Lab [84] y recientemente *ki-elements* [85], las cuales ya ofrecen soluciones para la evaluación cognitiva a través del análisis automático del habla. Algunos como Winterlights Lab incluso han desarrollado un robot de 61 cm de altura para este fin [86].

En esta tesis hemos localizado 91 publicaciones científicas, las cuales han sido clasificadas por año de publicación y por el tipo de extracción de características utilizada para el procesamiento de voz. A partir de la Figura 2-1 puede observarse un incremento exponencial en el número de trabajos publicados al respecto.

En los siguientes apartados hemos procedido a clasificar estas investigaciones atendiendo al tipo de medidas extraídas durante el proceso de extracción de características en convencionales y no convencionales. Cabe aclarar que, para analizar

los diferentes métodos de extracción de características, decidimos utilizar una clasificación convencional y no convencional en lugar de la lineal y no lineal. Esta última clasificación podría ser confusa ya que existen técnicas que son matemáticamente no lineales, pero describen fenómenos lineales. En esta tesis, nos referiremos a estas últimas como características convencionales, independientemente de la técnica matemática utilizada y, por su parte, las características no convencionales serán, en general, las que caracterizan los fenómenos de voz no lineales.

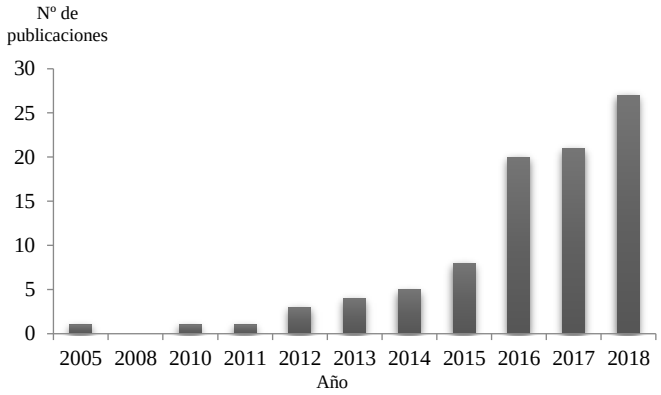


Figura 2-1 Artículos publicados sobre procesamiento automático de voz y habla aplicados a la detección de Alzheimer desde 2005 hasta principios de 2018 (en base a las publicaciones localizadas [36]).

2.2 Extracción de características convencionales

A partir de las publicaciones identificadas, hemos podido recoger los principales parámetros convencionales que han sido empleados para describir las características acústicas de la voz en la detección de AD. Estos parámetros se relacionan en la Tabla 2-1 [81], [87]–[91].

Las características que hemos definido como convencionales describen diferentes aspectos de la voz y de las emisiones verbales. Sin embargo, además de ellas, existe una serie de elementos que acompañan a las emisiones propiamente lingüísticas y que constituyen señales e indicios pero que, normalmente, son no verbales. Sugieren interpretaciones particulares de la información propiamente lingüística y es lo que se denomina rasgos paralingüísticos, los cuales han sido utilizados para obtener información del paciente mediante estadísticos de los primeros formantes, la concentración de energía en diferentes bandas de frecuencia, los coeficientes de Predicción Lineal (LPC) o los *Mel-Frequency Cepstral Components* (MFCC), entre otros [92].

De la Tabla 2-1 puede extraerse que la mayoría de los estudios publicados se desarrollan en base a diferentes características lineales simples, especialmente las centradas en características vocales y prosódicas [93], [94], [103]–[110], [95]–[102]. De manera general, estos estudios tienen en común que utilizan parámetros individuales para la evaluación del habla. Sus resultados han sido muy variados alcanzando precisiones de entre el 70-90%. En cualquier caso, han podido proporcionar evidencias sobre el uso potencial de tareas habladas simples para la evaluación automática de la

demencia muy temprana y su progresión, así como demostrar que la tecnología permite la detección precoz, automática, económica, remota y de gran escala [109], [110]. Algunos incluso han incluido en sus trabajos avatares informáticos [99], [100].

Tabla 2-1 Principales características convencionales utilizadas en la detección de AD a partir de voz [81], [87]–[91].

Aspectos temporales	**Duración**	Duración total Tiempo de fonación Velocidad de articulación y ritmo Alargamientos de tiempo
	Interrupciones	Porcentaje y número de pausas de voz Porcentaje sin voz Número y porcentaje de paradas de voz (MEDIA, MAX, MIN, %)
	Períodos sonoros	Número de pulsos analizados como voz Período (MEDIA, SD, Número medio de períodos de voz)
Aspectos frecuenciales	**Frecuencia fundamental**	F0 (MEDIA, MAX, MIN, SD, relF0SD, VR, relF0VR) SEO (SD, VR, relSEOVR) TEO(SD, VR, relTEOVR) *Short time energy* o *spectral centroid* *High and low global pitch* Autocorrelación
	Fluctuaciones	*Jitter* (*Short term, cycle-to-cycle,* perturbaciones en la frecuencia fundamental): *local jitter, local absolute jitter, relative average perturbation jitter*, PPQ.
Intensidad	**Amplitud**	Intensidad de señales sonoras y sordas (SD, MAX, MIN) SEO (SD) TEO (SD) RMSA
	Estabilidad fonatoria	*Period perturbation, shimmer (short term, cycle-to-cycle, perturbations in the amplitude of the voice): local shimmer, amplitude perturbation quotient* (APQ).
Calidad de la voz	**Ruido**	HNR NHR
Aspectos biomecánicos	***Vocal fold body Movement***	*Stiffness*
	Movimiento de la lengua	Uso de información de frecuencia como localización de los formantes y anchos de banda. Aunque son ampliamente conocidas, el uso generalizado de estas características en el análisis del habla en la AD aún no está tan extendido como en otras patologías, como puede ser la enfermedad de Parkinson.

SD: desviación estándar, VR: rango de variación, SEO: *Square Energy Operator*, TEO: *Teager-kaiser Energy Operator,* HNR: relación señal a ruido, NHR: relación ruido a señal, F0: frecuencia fundamental, RMSA: *Root Mean Square Amplitude.*

Cada estudio analizado ha tratado de arrojar luz sobre diferentes aspectos. Es el caso de un estudio basado en una técnica lineal de regresión paso a paso que, mediante las variables amplitud, frecuencia y periodicidad, demostró que el porcentaje de segmentos sordos explica una parte significativa de la varianza de la puntuación global obtenida en varias pruebas neuropsicológicas [94]. Hay estudios que se han centrado en demostrar que los resultados cambian atendiendo a la tarea concreta que se desarrolle. Uno de estos estudios [102] implementa una aplicación móvil que analiza las características vocales de los sujetos y demuestra mejores resultados de precisión para las tareas de habla espontánea y fluidez a la hora de discriminar entre sujetos de control

(en adelante, HC), sujetos con MCI y pacientes con AD [102]. En la misma línea, pero con tareas combinadas (habla espontánea, en adelante SS, y recuentos), otros estudios basados en el análisis de características vocales alcanzan precisiones en torno al 89% [103].

Existen bastantes trabajos que evalúan conjuntamente el proceso de extracción de características de la señal mediante diferentes parámetros vocales y la aplicación de diferentes clasificadores [101], [104], [108]–[110]. También han sido utilizadas técnicas de regularización para superar escasez de datos [109], [110]. Mediante el uso del Test de Mann-Whitney para la selección de características vocales, y el uso de modelos SVM y Naive Bayes (NB, en adelante), este mismo estudio obtuvo resultados que logran discriminar entre HC-MCI, 79% ± 5%; entre HC-AD, 87% ± 3%; y entre MCI-AD, 80% ± 5% (AD:26, MCI:23, HC: 15).

La eficiencia de las diferentes características lineales y la mejor selección de ellas han sido evaluadas en un estudio a partir de tareas de lectura en francés (AD: 14 MCI: 14 HC: 14) [105]. Las características analizadas fueron 4 relacionadas con el *pitch*, 2 con segmentos del habla y 9 con *jitter* y *shimmer*. Se emplearon dos clasificadores, kNN y SVM multiclase. El clasificador kNN con diez características se comportó mejor que la SVM en el grupo de sujetos de control (57% frente a 43%). Se encontraron, asimismo, altas tasas de confusión en la discriminación de perfiles MCI. Como conclusión se obtuvo que la precisión mejora con la reducción del tamaño de características. En la misma línea, un estudio [111] realizó sobre la base de datos DementiaBank un proceso de selección de características sobre varias medidas definidas en el mismo (Fracción de fragmentos localmente sordos, MFCC2, Curtosis -MFCC30, Media -MFCC30, Skewness -MFCC2, Media -MFCC16 o número de segmentos sordos, entre otros). Se aplicaron 4 clasificadores: NB, Random Forest (en adelante, RF), AdaBoost M1 (en adelante, AB) y Meta Bagging (en adelante, MB) [112] para ejecutar los experimentos con validación cruzada *k-fold*. Se logró una precisión de clasificación 94,7% con niveles de sensibilidad y especificidad de 97% y 91%, respectivamente, utilizando el clasificador NB, preprocesado para mitigar el ruido de fondo y las 20 mejores características. En la misma línea, otros estudios [101], además de realizar el análisis lineal en base a la significancia estadística de las características (valor de *p-value* < 0,05), lo han aplicado igualmente a las tareas realizadas (prueba Kolmogorov-Smirnov no paramétrico *p-value* < 0,10, para cada tarea). Se probaron como clasificadores automáticos kNN con 3-vecinos, Regresión Logística y NN. Los mejores resultados los consigue el clasificador NN con una precisión de 0,767.

El análisis desde una perspectiva meramente silábica también ha sido llevado a cabo en otros estudios (AD: 45, HC: 82) [106], aplicando rangos de medidas rítmicas, clínicas y métricas, basándose esta última en la desviación estándar de la duración de los intervalos silábicos. Este trabajo consiguió discriminar pacientes con AD y adultos mayores sin demencia con una precisión del 87%, especificidad 81,7% y sensibilidad 82,2%.

Por otra parte, aunque menos, también pueden encontrarse estudios que han basado su sistema únicamente en características paralingüísticas [113]. El interés de este estudio radica en que es independiente del idioma y, en el caso de este estudio, alcanza

una precisión en torno al 73%, si bien está más extendido en el campo la combinación de características paralingüísticas y lingüísticas [114].

También se han definido conceptos más amplios basados en combinaciones de los anteriores parámetros. Es el caso de *Automatic Spontaneous Speech Analysis* (en adelante, ASSA) [88], [90] que hace uso de tres familias: duración (segmentos de voz y sordos), *short-time energy* en el dominio del tiempo y el *spectral centroid* en el dominio de la frecuencia. Ofrece información útil para el análisis del habla espontánea y consigue una precisión del 75,2% mediante un clasificador SVM aplicado a la base de datos AZTIAHO [115].

Por su parte, el reconocimiento automático de emociones a partir del habla resulta de especial interés debido a que, como se expuso en el apartado 1.1.2, los pacientes de AD presentan cambios en la manera de manifestar sus emociones con respecto a los sujetos cognitivamente sanos. Actualmente no hay acuerdo sobre el número de emociones a analizar [116]. La mayoría de las investigaciones se centran en cuatro emociones básicas: ira, miedo, tristeza y felicidad [117] o, en algunos casos, también incluyen sorpresa y asco [118]. Otras investigaciones se centran en el desarrollo de aplicaciones en tiempo real y enfatizan la utilidad de representar las emociones en un plano de evaluación en términos de dos o más niveles o dimensiones continuas [119]. En la práctica, los niveles de activación y valencia son las dos dimensiones más utilizadas [92]. La activación está relacionada con la intensidad percibida de la emoción y el nivel de valencia tiene que ver con la sensación placentera percibida de ese estímulo [92]. En la Figura 2-2 se representan ambos planos y sus emociones asociadas.

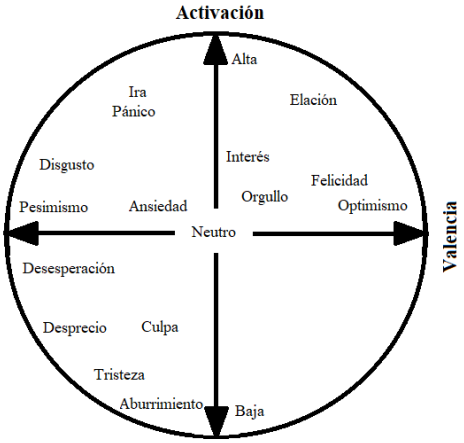


Figura 2-2 Clasificación de las emocionales: planos de activación y valencia [92].

Investigaciones previas en el reconocimiento automático emociones a partir del habla presentan un proceso de extracción de características basado en la caracterización de aspectos prosódicos, como el contorno de tono, el contorno de energía, la duración de las fonaciones [120]–[122] o el *Teager Energy Operator* [123] que están relacionados con la estructura gramática y el estrés léxico. Los aspectos paralingüísticos se reflejan en características como el primer formante [121], [122] o la concentración de energía en diferentes bandas de energía [124]. El enfoque clásico en el reconocimiento de las emociones a partir del habla utiliza un amplio conjunto de características, entre 13 y 1941 características paralingüísticas y prosódicas [124]. Sin embargo, este rango de

características no es adecuado para aplicaciones en tiempo real, debido a la complejidad de algunas de las características y al tiempo de procesamiento.

Los análisis emocionales del habla, *Emotional Speech Analysis* (ESA), aplicados a la detección de AD que hemos localizado se basan en otras tres familias de parámetros, a saber, características acústicas (pitch, SD, máximos y mínimos del pitch, SD, máximos y mínimos de la intensidad, SD del período entre otros), características de la calidad de la voz (*shimmer, jitter* local, HNR, NHR, autocorrelación) y características de la duración (fragmentos sonoros o sordos). El *short-term energy* es la principal característica analizada.

Por su parte, el concepto Temperatura Emocional (TE), conjuga varias características prosódicas y paralingüísticas a partir de la segmentación temporal de la señal del habla. De cada fragmento de señal se extraen dos características prosódicas y 4 paralingüísticas, relacionadas con el pitch y la energía, respectivamente [92]. Mediante la TE y el uso de una SVM es posible clasificar las muestras como patológicas y no patológicas obteniendo muy buenos resultados para la detección específica de la AD [92]. Conjuntando análisis ASSA y TE, se ha podido demostrar una pérdida significativa de fluidez en personas con AD con respecto al porcentaje y duración de los segmentos sonoros/sordos [88], [125] y una precisión del 92.24%. Combinando ASSA, ESA y TE, los resultados de precisión pueden alcanzar, mediante el uso de SVM, en torno al 94% [90].

Otras corrientes interesantes en el procesado de voz aplicado a la detección de AD hacen uso de la información contenida en los golpes de audio y silencio, para lo que se utiliza el *Voice Activity Detector* (en adelante, VAD). A partir del VAD el discurso se distingue del ruido por una comparación espectral de una señal de entrada con una estimación almacenada del ruido [126]. Por su parte técnicas más complejas como el Word Spotting (WS) se basan en agrupar imágenes de la palabra en *clusters* de palabras similares basándose en similitudes [127]. Como en los mencionados anteriormente, las características analizadas más ampliamente siguen siendo las relacionadas con los segmentos de voz o silencio [128] y características prosódicas en general [129], [130]. Sin embargo, estas técnicas aplicadas a la detección precoz de enfermedades como la AD permiten, una vez realizada la transcripción de la señal de voz a texto, llevar a cabo la evaluación lingüística además de la acústica. Dejan abierta la posibilidad de realizar exámenes automáticos que se centren en la estructura sintáctica, léxica [131], semántica [132], fonética [133]–[135] o una combinación de todos ellos [136], [137].

Algunos trabajos [138] han expuesto la alta discriminación que pueden alcanzar las características prosódicas (siendo muy útil la relación de silencio), por ejemplo, frente a otras de carácter lingüístico (como relaciones nombre-pronombre). Otros sistemas estudiados [132] concluyeron que las características que ayudan a distinguir MCI de HC se basan principalmente en características morfológicas (número de palabras, signos de puntuación, sustantivos…). En este trabajo, concretamente, se extrajeron características lineales como el tiempo hablado y de silencio, dubitaciones, morfológicas y semánticas (relacionadas con la memoria). Se usó un clasificador SVM y, debido al pequeño tamaño del conjunto de datos, se aplicó validación cruzada y

etiquetado mayoritario. Usando todas las características, el sistema alcanzó una tasa de éxito del 69,1%.

Varios estudios han demostrado peores resultados cuando se emplea un gran número de características en el análisis [139]. En un trabajo realizado en 2016, mediante tareas de habla espontánea en alemán (sujetos de control: 80, sujetos con deterioro cognitivo asociado al envejecimiento (en adelante, AACD): 13 y pacientes con AD: 5) la mayoría de las características prosódicas se crearon a partir de segmentos de silencio y segmentos de voz obtenidos del VAD, el cual utiliza un reconocedor Modelo Hidden-Markov. La tasa de éxito del clasificador LDA utilizado fue de 80,4% utilizando 14 funciones lineales [128], si bien en una publicación posterior [129] alcanzan una tasa de éxito general del 85,7% haciendo uso solo de diez características.

Se han desarrollado también trabajos basados en técnicas *n-gram* [133], estudiando la correlación de la frecuencia de diferentes partes del habla y DAT en la base de datos ACADIE. La tasa de éxito alcanzada es alta (95% en AD vs HC y tasa de éxito del 70% para discriminar dos clases de demencia). También basado en modelos *n-gram* [135] otro estudio ha analizado la correlación entre los cálculos estadísticos y el MMSE. Empleó como base de datos DementiaBank y un clasificador LDA con validación cruzada. En este caso, el 80,4% de las muestras fueron clasificadas correctamente.

Varios estudios han trabajado para evidenciar la eficacia de sistemas automáticos aplicados a las transcripciones de voz frente a aquéllos manuales [131], [140]. Se alcanzaron puntuaciones cercanas a los del set de características manuales y los mejores resultados fueron para los clasificadores SVM (90,6% y 88%, respectivamente). Otro estudio en la misma línea, para demostrar la eficacia de los métodos automáticos [108], fue capaz de separar las dos clases con una puntuación F1 de 78,8% (MCI: 48, HC: 38, tarea: habla espontánea). Sin embargo, en este caso, el clasificador RF funcionó ligeramente mejor que el SVM con el conjunto de características automáticas. Las diferencias más significativas entre los dos grupos se encontraron en la duración del habla en la tarea de recuerdo diferido y en el número de pausas para la tarea de responder preguntas. Sin embargo, no se pudo extraer una conclusión definitiva con respecto a qué algoritmo de aprendizaje automático - SVM o RF- es más adecuado para la tarea de clasificación dada.

Un estudio reciente [141] se ha centrado en la aplicación de la Teoría de la Estructura Retórica (RST) [142], [143]. A partir de la base de datos DementiaBank y Carolina Conversations Collection se extraen automáticamente relaciones del discurso de acuerdo con la RST a partir de tareas de habla espontánea. Extraer las relaciones RST segmentando el texto original, construyendo árboles a partir de esos segmentos y, posteriormente, seleccionando las características más relevantes. Las relaciones discursivas, especialmente las que involucran la elaboración y atribución, son indicadores significativos de AD en el habla. En las transcripciones, por el contrario, implican relaciones de contingencia lógica. Se observan diferencias significativas en la estructura del discurso entre personas con AD y HC en transcripciones de habla espontánea. También se observa, a partir de la base de datos DementiaBank, que existen diferencias sutiles incluso entre subtipos específicos de demencia.

En esta misma línea, existen estudios [144] que, a partir de las denominadas unidades de contenido de información (ICU), han podido demostrar que los grupos AD tienen una densidad de ideas significativamente menor que los HC en todos los escenarios, especialmente en tareas de descripción de imagen. Otro trabajo [145] se basa en el concepto densidad de ideas proposicionales (PID) y desarrolla DEPID, método para el cálculo de PID, y su versión DEPID-R que permite excluir la repetición de ideas. Se han desarrollado también entornos computacionales unificados para análisis automatizado de la producción de lenguaje en entornos clínicos como es Coh-Metrix-Dementia [146]–[148]. El sistema se construyó sobre varias herramientas basadas en procesamiento del lenguaje natural (NLP, por sus siglas en inglés). A partir de una muestra de texto transcrito producido por un sujeto en portugués se extrajo 73 métricas textuales. En una primera fase demostró tener una alta precisión de clasificación [146]. En un estudio posterior de 2016 [147] se concluyó que para entrenar los modelos de regresión y clasificación, se debe reducir el gran conjunto de características en Coh-Metrix-Dementia y se compararon tres métodos de selección de características. En 2017 [148] se analizaron características lineales como la *informatividad*, el número de proposiciones de cada texto; coherencia global, emisiones vacías, densidad de ideas total, características semánticas y cantidad de *modalizaciones*. Se utilizó una prueba no paramétrica Kruskal-Wallis para comparar el rendimiento entre los tres grupos (60 personas. mAD, MCI y HC) con respecto a las variables. Los sujetos con AD leve mostraron menor discurso informativo y estructura narrativa y un mayor deterioro en la coherencia global y *modalización*. Sin embargo, no fue posible discriminar entre MCI y HC.

Otros estudios [114] han combinado modelos de efectos mixtos lineales (base de datos WRAP [149]). Las transcripciones se codifican para análisis automáticos mediante *Computer Language Analysis* (CLAN) [150], que incluyen códigos para pausas rellenas y sin completar, repeticiones, revisiones, unidades semánticas, errores (semánticos, fonológicos, léxicos) y comportamientos no verbales. Las unidades semánticas, partes del discurso, relaciones gramaticales y otros cuantificadores son extraídos automáticamente por CLAN utilizando MOR y MEGRASP [151]. Los participantes con MCI temprana (eMCI) disminuyen más rápidamente en características de fluidez del habla y contenido semántico que los HC. Sin embargo, las medidas de diversidad léxica y complejidad gramatical no se asocian con el estado eMCI.

2.3 Extracción de características no convencionales

En estados más avanzados de la AD, la voz se vuelve aperiódica, ruidosa, irregular y caótica. En ocasiones estas características limitan la capacidad de los parámetros convenciones para obtener información útil sobre la patología. En base a esta realidad, los investigadores han desarrollado nuevos parámetros complejos y más robustos. Comparados con los convencionales, éstos ofrecen una información más precisa, a pesar de ser clínicamente menos interpretables [81]. Al respecto, se ha podido demostrar que las señales de voz patológicas pueden ser descritas por análisis dinámico no lineal

[152]. En la Tabla 2-2 se recogen los principales parámetros clasificados como no convencionales aplicados al análisis del habla en la detección de la AD.

Tabla 2-2 Principales características no convencionales utilizadas para la detección de AD.

Características no lineales	**Descripción**
Dimensión correlación (CD)	Geometría atractor en el espacio fase [153]–[155].
Dimensión fractal (FD)	Número de bloques construidos básicos que forma un patrón [153], [156].
Complejidad Ziv-Lempel (ZL)	Cuantifica la regularidad de una serie temporal [157].
Exponente de Hurst (HE)	Describe posibles dependencias *long-term* en la señal de habla [158].
Entropía	
Entropía de Shannon (SHE)	Medida de incertidumbre, cuantifica la complejidad de un sistema.
Second-order Rényi Entropy (RE)	RE cuantifica la pérdida de información en tiempo en un Sistema dinámico [159], [160].
Correlation Entropy (CE)	
First-order And Second-order Rényi Block Entropy (RBE1 y RBE2)	Indicación CE de la predictibilidad de las series de tiempo no lineales [159], [160].
Entropía aproximada (AE)	La única diferencia entre AE y SE [161] es que SE no evalúa una autocomparación de vectores embebidos [153], [161], [162].
Sample Entropy (SE)	
Recurrence Probability Density Entropy (RPDE)	Incertidumbre en la medida del período de pitch [163].
Primer Mínimo de Información Mutua (FMMI)	Discrimina entre las diferentes calidades de la voz [159].
Largest Lyapunov Exponent (LLE).	Incluye una medida de sensibilidad a una condición inicial [153], [158].
Detrended Fluctuation Analysis (DFA)	Caracteriza la autosimilitud del gráfico de la señal desde un proceso estocástico [163], [164].

Existe una variedad de trabajos en torno al análisis de la voz que emplean características no convencionales de la señal, aunque son relativamente recientes. Entre otros, han sido explorados los espectros de orden superior y el acoplamiento de fase cuadrático de las señales de habla espontánea para grupos HC y AD usando *bispectrum* y *bicoherencia*. Se ha podido comprobar que los acoplamientos de la fase cuadrática de la señal procedente del habla espontánea de personas con AD se reducen en comparación con sujetos HC. Los armónicos acoplados en la fase del habla se desplazan a las frecuencias más altas en sujetos AD. En 2016 uno de estos estudios, a partir de la variabilidad de la frecuencia y la dinámica de la señal del SS de grupos AD y HC [165], examinó la superficie del error cuadrático medio (MSE) y la cuantificación de las curvas de nivel de ambos grupos. Concluyó que las señales de voz transitan desde un estado caótico de alta dimensión en sujetos HC a un movimiento caótico de baja dimensión en pacientes con AD. Estos estudios, generalmente, basan su potencial en la combinación con características lineales. Un claro ejemplo de ello es la aplicación de las técnicas ASSA y *Emotional Response Analysis* (en adelante, ERA) con la Dimensión Fractal de Higuchi (HFD) y una serie de características acústicas y estacionarias [166]. Se aplicaron implementando un clasificador de red neuronal MLP con resultados muy satisfactorios [167], [168].

Las oscilaciones cuasiestáticas de las cuerdas vocales y el ajuste proporcionado por el tracto vocal son procesos no lineales producidos principalmente cuando se pronuncian las consonantes y podrían describirse mediante fractales. Según Mandelbrot [169], fractal es un patrón geométrico que se itera en escalas más pequeñas o más grandes para producir formas o superficies similares auto-similares, irregularidades que

la geometría euclidiana no puede representar. La forma de onda de la señal del habla muestra concretamente tanto periodicidad como autosimilaridad según las vocales y las consonantes que forman las sílabas juntas. Algunos estudios ya han realizado análisis fractal sobre la señal de voz [170].

Otros afirman que la información adicional proporcionada por las características no lineales a partir de dimensiones fractales (en adelante, FD) podría ser especialmente útil cuando los datos con que se cuenta son escasos [171]. En esta línea, ciertos estudios exploraron la problemática de los conjuntos de datos desequilibrados mediante clasificadores de una sola clase y se concluyó que el uso de FD mejora el rendimiento del sistema para estos casos [89], [168]. En general, el uso de características de FD con clasificadores multiclases de una y dos clases han demostrado obtener resultados muy satisfactorios [172]. En trabajos posteriores los autores [166], [173], profundizaron en la utilidad del uso de FD para mejorar las técnicas de modelado basadas en las características de respuesta emocional. Para ello, a partir de la base de datos AZTIAHO, se propuso usar algoritmos de Higuchi, Katz y Castiglioni y agregar estas nuevas características no lineales al conjunto que alimenta el proceso de entrenamiento [173]. Los mejores resultados se obtuvieron con una combinación de características acústicas, de calidad y duración, con TE y con un set que englobaba la dimensión fractal (HFD, KFD y CFD) y sus variaciones estadísticas para la señal completa. También se analizaron diferentes clasificadores obteniendo los mejores resultados con un clasificador MLP. La precisión del sistema final superó para todos los grupos el 90%.

Investigaciones recientes con sistemas complejos muestran que las geometrías y fenómenos que evolucionan naturalmente no se pueden caracterizar por una sola relación de escala (sistema monofractal), ya que las diferentes partes del sistema tienen una escala diferente, y que solo el análisis multifractal podría medir su dinámica interna con mayor precisión [174]–[176]. Los multifractales son objetos auto-similares más complicados que consisten en fractales ponderados de manera diferente con diferentes dimensiones no integradas y sus propiedades de escala varían en cada una de las regiones de los sistemas [177]. La naturaleza multifractal del habla ha sido analizada para la representación y caracterización de algunas obras [178].

Detrended Fluctuation Analysis (en adelante, DFA) es un método de análisis de escalamiento donde el exponente de escala (similar a un exponente de Hurst de una escala o FD) se utiliza para cuantificar la correlación de largo alcance de la señal estacionaria y no estacionaria [179]. Otro método, el *Multifractal Disactivated Fluctuation Analysis* (en adelante, MFDFA) ha sido aplicado para estudiar el comportamiento de escalamiento multifractal de diversas series de tiempo invariantes de escala no estacionaria por Kantelhardt [180]. Algunas fuentes [176], afirman que los resultados obtenidos por este método fueron más fiables en comparación con métodos como el análisis de onda, transformada wavelet discreta, módulo de transformada de wavelet máxima, promedio móvil descendente, promedio móvil de banda, Análisis de Fluctuación Anulada Modificada, etc. [170], [180]–[182]. En este mismo estudio [176], se definió una nueva característica para la detección de emociones a partir del habla utilizando el método MFDFA aplicado a la AD. Se propuso un parámetro cuantitativo para categorizar varias emociones mediante el análisis de los detalles no estacionarios

de la dinámica de la señal del habla. Para ello se empleó la base de datos TESS [183]. El exponente Hurst y el ancho del espectro multifractal se calculó para 1200 clips (200 para cada una de 6 emociones) utilizando el *software* de Matlab y según el método prescrito por Kantelhardt [180]. Los resultados arrojaron luz, a partir de la tendencia de los valores del exponente de Hurst y los anchos del espectro multifractal, sobre la posibilidad de distinguir claramente entre las emociones fundamentales y esto, a su vez, aplicarlo a la detección precoz de la AD [176].

2.4 Deep Learning

En los últimos tiempos se ha hecho notable la tendencia al uso de modelos de clasificación de datos basados en el Aprendizaje Profundo o *Deep Learning* (DL, por sus siglas en inglés) [184]. Estos modelos son usados para capturar interacciones complejas de características no lineales en datos multimodales como pueden ser características de audio y vídeo.

Dentro del *Machine Learning* (ML), el DL permite implementar modelos computacionales que están compuestos por múltiples capas de procesamiento para aprender representaciones de datos con múltiples niveles de abstracción. Estos métodos han mejorado drásticamente el estado de la técnica, muy especialmente en reconocimiento y procesado de voz [185]. El aprendizaje profundo descubre una estructura compleja a partir de ingentes conjuntos de datos mediante el uso de algoritmos de retropropagación, los cuales indican cómo una máquina debe cambiar los parámetros internos que utiliza para calcular la representación en cada capa, a partir de la representación en la capa anterior. Las Redes Convolucionales Profundas han producido adelantos en el procesamiento de imágenes, video, voz y audio, mientras que las Redes Recurrentes han resultado muy interesantes con datos secuenciales, como el texto y el habla [185]. Como se puede ver, diferentes arquitecturas se prestan para resolver diferentes tipos de problemas.

En un estudio de 2017, un sistema automático [130] basado exclusivamente en parámetros lineales procesó cada muestra para separar las pausas de la voz mediante técnicas VAD. Con las muestras del habla divididas en pausas y eventos del habla, calculó 22 métricas, incluyendo medidas de duración total del discurso, número de pausas y fracción del tiempo de conversación pausada, tono y energía, entre otras. Se configuró una Red Neuronal Recurrente o *Recurrent Neural Network* (en adelante, RNN) con 2 capas ocultas, cada una con 256 neuronas y alcanzó una tasa de éxito del 94% en la discriminación entre sujetos AD y HC. Incluyó una combinación de características acústicas del habla, características lingüísticas extraídas de la transcripción automática del discurso, así como la puntuación y resultados del test MMSE.

También en 2017, se desarrolló la plataforma ALZUMERIC [186] donde especialistas médicos pudieron recopilar muestras de voz y el sistema extrajo automáticamente características de las señales para ofrecer un principio de diagnóstico en base a la articulación, calidad del habla, análisis de la respuesta emocional,

percepción del lenguaje y dinámica compleja del sistema. Se realizó una selección de características clásicas, perceptuales y no lineales mediante la prueba Mann-Whitney U-test con un valor de p <0,1. Se utilizaron, asimismo, cuatro clasificadores: k-NN, SVM, MLP y una Red Neuronal Convolucional o *Convolutional Neural Network* (en adelante, CNN). El programa WEKA [112] se eligió para llevar a cabo los experimentos. Para el entrenamiento y validación se utilizó validación cruzada k-*fold* con k = 10 [187]. La tarea que mayor precisión obtuvo fue la de habla espontánea con una tasa de éxito del 95%. Los mejores resultados, en base a la selección de las características más relevantes por pruebas estadísticas (bajo criterios médicos) y métodos de clasificación, se obtuvieron con Mann-Whitney U-test y SVM [186].

En una continuación del trabajo [188], de nuevo, se incluye el aprendizaje profundo haciendo uso de una CNN y modelado multirreal no lineal. Se incluyen características lineales (HNR, *pitch, jitter*), perceptuales (MFCC...) y no lineales (fractales, entropía de Shannon, entre otros) [188]. Finalmente, se pudo concluir que la integración automática de las características más relevantes usando CNN (2 capas y 20 neuronas) proporciona información útil no disponible en otras pruebas estadísticas.

En 2018, uno de los más recientes estudios que hemos podido localizar [189], analizó tres tareas con diferentes niveles de complejidad del lenguaje y la clasificación automática se llevó a cabo utilizando MLP y CNN. De nuevo, se hizo uso de características lineales clásicas, perceptivas, dimensión fractal Castiglioni y Entropía de Permutación Multiescalar. Por último, concluyó que las características más relevantes se seleccionan mediante la prueba U no paramétrica de Mann-Whitney.

Otro estudio [113], propuso un método de detección automática de la AD utilizando una Red Neuronal Convolucional Cerrada o *Gated Convolutional Neural Network* (en adelante, GCNN), la cual se podía entrenar con una cantidad relativamente pequeña de datos y capturar la información temporal en funciones paralingüísticas de audio. Además, como no utilizó ninguna característica lingüística se podría, asimismo, aplicar fácilmente a cualquier idioma. Usó el *software* openSMILE [190] y una serie de características expuestas en la conferencia INTERSPEECH 2009-2012 [191]–[193]. El método propuesto alcanzó una precisión del 73,6%. En cualquier caso, en su momento, quedó pendiente evaluar el enfoque en diferentes idiomas.

De manera general, a día de hoy estos trabajos son sólo algunos de los estudios desarrollados en torno al procesado automático del habla aplicado a la detección de AD, sin embargo, cada vez más investigaciones incluyen este tipo de técnicas de Aprendizaje Profundo en busca de soluciones consistentes que puedan dar respuesta a un proceso tan complejo como es el comunicativo.

Capítulo 3 Metodología del estudio

En este capítulo se describe la metodología empleada para validar la hipótesis presentada en el Capítulo 1, donde se exponía que es posible caracterizar la voz de pacientes de Alzheimer y discriminarla frente a la de sujetos sanos a partir de las muestras recogidas por un sistema conversacional automático.

En la Figura 3-1 se presenta la metodología empleada en esta tesis mediante un diagrama de bloques donde se representan las diferentes muestras recogidas, los entrevistadores utilizados y las características de la voz extraídas, así como los diferentes análisis realizados y la posterior extracción de conclusiones.

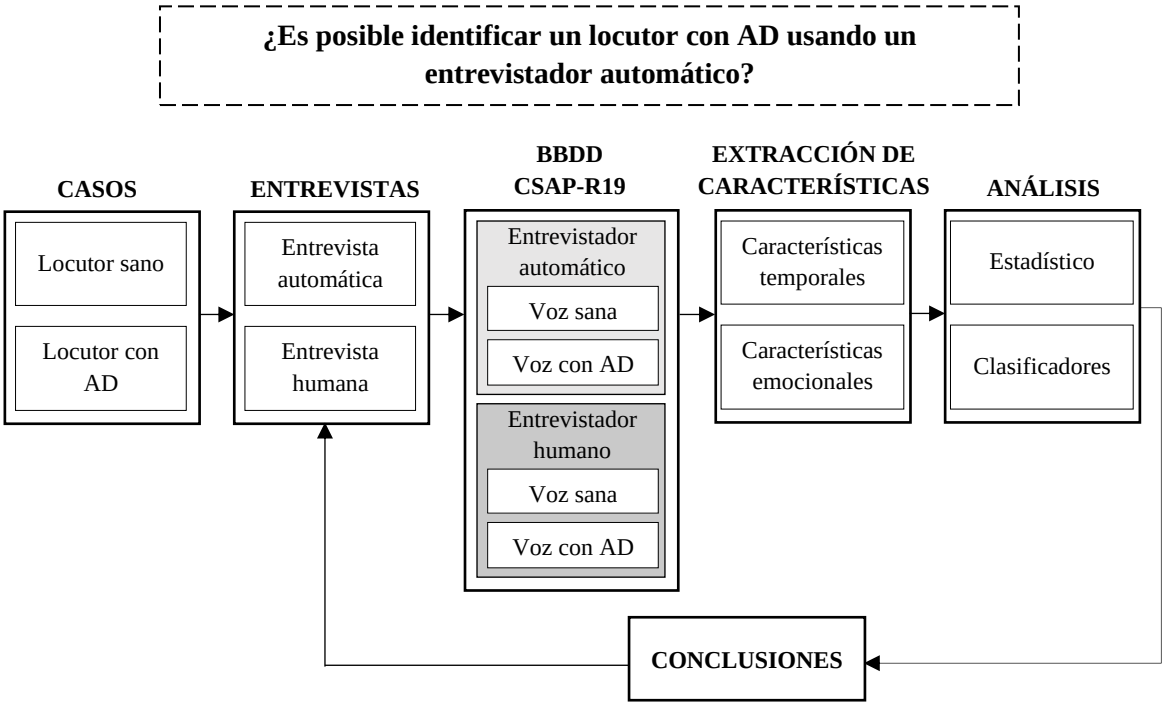


Figura 3-1 Diagrama de bloques: metodología empleada en la tesis.

Para obtener las muestras indicadas en la Figura 3-1 se ha creado una base de datos adecuada a partir de una herramienta automática de grabación (*software* Prognosis) y entrevistas manuales. Tal y como se indica en el esquema anterior, se ha realizado un proceso de extracción de características temporales y de carga emocional y, por último, se han analizado las mismas mediante una serie de estudios estadísticos y diferentes modelos de clasificación con el fin de profundizar sobre la viabilidad de nuestra hipótesis.

3.1 Base de datos

Para la creación de la base de datos CSAP-R19 utilizada en esta tesis se han tenido en cuenta y recogido dos tipos de muestras: las obtenidas a partir de un entrevistador humano y las obtenidas a partir de un entrevistador automático. En adelante, ambos tipos de muestras recibirán el nombre de muestras de habla espontánea y muestras de habla inducida, respectivamente. Una característica importante de las muestras que conforman la base de datos CSAP-R19 es que han sido recogidas de manera puntual, no longitudinal en el tiempo. Todas las muestras para un mismo participante se tomaron en una única sesión.

Yendo a los detalles de las muestras, de cada sujeto se ha obtenido en total 4 grabaciones, de las que tres son de habla inducida y una de habla espontánea. La duración media de todas las grabaciones es de 34,5 segundos, a una frecuencia de muestreo de 44.100 Hz. El tipo de archivo de grabación utilizado ha sido WAV.

De manera general, para el proceso de grabación se ha usado como herramienta principal un ordenador (Intel Core i7, 6 GB RAM. 750 Gb HDD. Monitor 16.9” LCD) y unos auriculares con micrófono Ozone Rage ST. Auricular: sonido estéreo, impedancia: 32 Ω, rango de frecuencias: 20~20 kHz, dimensiones: ø40mm, entrada máxima: 85 dB (At1.0 kHz). Micrófono (Impedancia: ≤ 2,2 kΩ, Sensibilidad: -38 dB (at 1 KHz), directividad: omnidireccional, frecuencia de Respuesta: 20 ~ 20 kHz, Dimensiones: 5,0 mm × Φ6,0).

Es interesante aclarar que las grabaciones se han realizado siempre en el entorno de los pacientes y de los sujetos de control, a fin de que el resultado sea lo más realista posible e influya menos en la grabación. Por su parte, cabe mencionar que todos los familiares han firmado previamente el consentimiento informado del estudio y han sido debidamente informados del certificado favorable emitido por el Comité Ético de la Universidad de Las Palmas de Gran Canaria (en adelante, ULPGC), con referencia CEIH-2014-01 y fecha 17 de julio de 2014.

Respecto al etiquetado de las muestras, se ha seguido una nomenclatura común en base a la estructura mostrada en la Figura 3-2. Así, por ejemplo, el fichero S2F1936_20160612_472248 realizado el 12 de junio de 2016 correspondería a una mujer con patología leve, nacida en 1936.

Como se ha comentado, la base de datos CSAP-R19 está compuesta de grabaciones de dos clases atendiendo al tipo de entrevistador que guía la conversación. Las grabaciones obtenidas a partir del entrevistador humano fueron realizadas por un

miembro del equipo de investigación utilizando un ordenador portátil donde se han almacenado las distintas grabaciones. Además, se ha utilizado el *software* Audacity® para realizar la grabación junto con los auriculares con micrófono Ozone Rage ST [194]. Al realizar la entrevista se pidió o animó a cada uno de los sujetos a hablar sobre lo que quisieran, sobre un tema libre, con el objetivo de lograr una grabación espontánea del habla. En este caso las muestras variaron entre 30 segundos y 2 minutos.

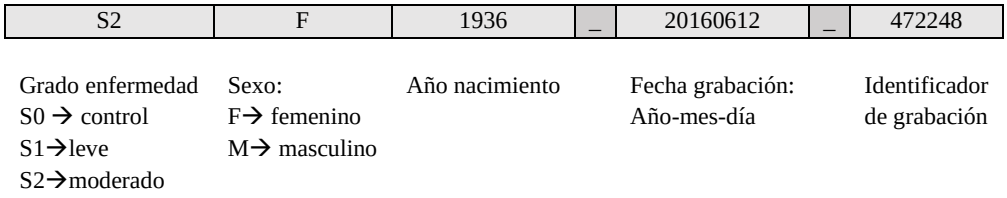


Figura 3-2 Estructura de etiquetado de muestras CSAP-R19.

Por su parte, las muestras obtenidas a partir del entrevistador automático se obtuvieron haciendo uso del *software* Prognosis, junto con el ordenador portátil y los auriculares con micrófono Ozone Rage ST [194] utilizados con el entrevistador humano. En total se obtuvieron 3 grabaciones de habla inducida por cada participante. Pueden consultarse más detalles sobre el *software* Prognosis en el apartado 4.1.

3.2 Extracción de características: medidas temporales

Para llevar a cabo la extracción de características ya sean de medidas temporales o, como veremos más adelante, de Temperatura Emocional, comenzamos realizando el preprocesado de las grabaciones. Este consiste en revisar cada una de las grabaciones para eliminar sonidos que puedan afectar al procesado posterior. Dichos sonidos incluyen los ruidos generados por objetos sonoros externos al locutor o el ruido de fondo producido por el micrófono o sala. En esta fase se ha utilizado el *software* Audacity®.

Una vez las grabaciones han sido debidamente acondicionadas, se lleva a cabo el procesado de las mismas por medio de diferentes algoritmos que nos permiten extraer las características temporales. A partir del *software* Matlab®, primeramente se ha hecho uso de un detector de actividad de voz o VAD [126]. El funcionamiento de este algoritmo se basa en seleccionar de la grabación los fragmentos de habla frente a los de silencio.

Con el habla identificado y seleccionado, a partir del vector generado por el VAD [126], se hacen mediciones de la duración del habla o silencio en una misma muestra. Las medidas obtenidas a partir de este vector han sido: tiempo medio de fragmentos sonoros, media aritmética, varianza, *skewness* y curtosis de las duraciones de habla, como se ha especificado, para cada muestra.

3.3 Extracción de características: Temperatura Emocional

A partir de las grabaciones preprocesadas y debidamente acondicionadas, se procede al procesado de las mismas mediante diferentes algoritmos para extraer características de Temperatura Emocional discreta y continua [92].

El algoritmo empleado ha sido propuesto por el Dr. Jesús B. Alonso Hernández y empleado y validado en numerosos artículos científicos. La función se basa en analizar y cuantificar para cada grabación las variaciones de activación de las emociones ofreciendo dos tipos de resultados, la Temperatura Emocional discreta (te_d) y la Temperatura Emocional continua (te_c). La primera es un número y da información general de toda la grabación, indica si el habla en su conjunto es emocional $te_d > 50$, o no. Para el cálculo de la segunda característica relacionada con la Temperatura Emocional, se segmenta la grabación en diferentes fragmentos y, de cada uno, se obtiene un nuevo valor de Temperatura Emocional dando como resultado final un vector. De este vector se obtiene, a su vez, el valor medio, la varianza, el *skewness* y la curtosis, que permitirán caracterizar la Temperatura Emocional continua.

3.4 Estudio estadístico

Una vez se han calculado las características temporales y de Temperatura Emocional de las grabaciones se realiza un análisis estadístico de las mismas empleando, para ello, las herramientas Stata® y Excel®. El estudio estadístico se ha realizado en tres bloques principales: medidas temporales, Temperatura Emocional y análisis conjunto.

De manera general, lo adecuado es que en el caso de que las muestran tengan una distribución normal se aplique un estudio paramétrico y que si, por el contrario, no siguen dicha distribución, se aplique un estudio no paramétrico. Al desconocer a priori este hecho, para comenzar el estudio, en Stata, se realiza un análisis paramétrico basado en una regresión lineal tanto de las características temporales, como de Temperatura Emocional. Posteriormente, y en base a los resultados obtenidos, se realiza un análisis de estadística descriptiva y un estudio no paramétrico basado en la Prueba de Suma de Rangos de Wilcoxon, en la Prueba de Kruskal Wallis, y en la Prueba de la Mediana.

Como resultado del análisis estadístico obtenemos diferentes tablas comparativas y *boxplots* de las características temporales y de Temperatura Emocional que permiten visualizar de forma gráfica la distribución de las muestras.

Para llevar a cabo los análisis mencionados, en primer lugar hemos creado un documento Excel y hemos incluido todas las variables a estudiar en diferentes columnas, junto con todos los datos asociados a cada una de las muestras (a saber: sexo, enfermedad, grado, año nacimiento, MediaHabla, VarHabla, SKWHabla, KRTHabla, IndHabla, TE, MediaTEc, VarTEc, SKWTEc, KTRTEc). Una vez importados los datos desde Excel a la aplicación de análisis estadístico, para observar de una manera sencilla diferencias entre las variables de las diferentes poblaciones bajo estudio, Stata ofrece el

comando *summarize*. Este comando nos permite disponer de una tabla con la media, la desviación estándar y el rango de las variables importadas (estadísticos descriptivos).

A partir de estos valores globales para cada variable, partiendo de la suposición inicial de que las muestras siguen una distribución normal, hemos realizado un estudio paramétrico mediante el método de regresión lineal y su posterior análisis de residuos. Este paso es fundamental para poder comprobar la adecuación del análisis paramétrico a nuestro estudio. El test de normalidad de los residuos realizado ha sido, concretamente, el test de *skewness* y curtosis donde, a partir de los valores de *Chi* cuadrado, podemos conocer si la regresión es correcta [195].

Teniendo en cuenta que la estadística no paramétrica es una rama de la inferencia estadística cuyos cálculos y procedimientos están fundamentados en distribuciones desconocidas, una vez realizado el análisis paramétrico, bajo la hipótesis de que las muestras no siguen una distribución normal, hemos llevado a cabo el estudio no paramétrico.

El primero de los test, la Prueba de Suma de Rangos de Wilcoxon, constituye la base para el resto de pruebas que utilizan rangos y, cuando se trabaja con variables medibles pertenecientes a una escala ordinal, como es este caso, se trata del test no paramétrico más indicado [196]. Como segundo test hemos realizado la Prueba de Kruskal Wallis, una generalización del procedimiento de la Prueba de Suma de Rangos de Wilcoxon [197]. Esta prueba nos permite comparar más de dos muestras con el propósito de conocer si proceden de la misma población o si hay diferencias entre las medidas de tendencia central de más de dos poblaciones. El último test no paramétrico realizado es la Prueba de la Mediana, a la que podemos considerar un caso especial de la prueba del *Chi* cuadrado. Esta prueba permite determinar si dos muestras independientes provienen de poblaciones con la misma mediana siempre que la variable esté, al menos, en escala ordinal [198].

Para cualquiera de los tres estudios no paramétricos, valores de $Prob|z|$ superiores a 0,05 ($Prob|z| > 0,05$), serán los casos en los que no hay diferencia al comparar las muestras de las diferentes poblaciones. Obtener valores de $Prob|z|$ por debajo de 0,05 (significa que existe menos del 5% de probabilidad de que dos poblaciones sean la misma) nos permitiría rechazar la hipótesis nula y, por tanto, que nos encontramos con dos poblaciones diferentes, capaces de ser segregadas a partir de la variable analizada.

Además de los análisis anteriores, también se ha realizado el análisis multivariante de la varianza MANOVA, centrado en las variables bajo estudio como un conjunto y no como características individuales. Concretamente se ha llevado a cabo para el set de características temporales, el de Temperatura Emocional y para el set completo de las diez características. En este caso, para realizarlo, hemos verificado previamente si las muestras siguen una distribución gaussiana y, por tanto, es conveniente realizar este tipo de análisis. La prueba de hipótesis utilizada se ha basado en medidas de asimetría.

3.5 Clasificadores

Para llevar a cabo la clasificación multivariante de las muestras obtenidas de uno y otro entrevistador se han empleado dos metodologías de trabajo diferentes. La primera, basada en el *software* de análisis estadístico ya mencionado anteriormente, Stata, y la segunda, haciendo uso del sistema de cómputo numérico Matlab, concretamente de la herramienta *Classification Learner App* ofrecida como aplicación por MathWorks®. Asimismo, se ha hecho uso de la herramienta Excel para almacenamiento, recopilación y gestión de datos y resultados.

A partir del programa Stata se han empleado los clasificadores LDA, logístico y kNN (para $n = 1$, $n = 3$, $n = 5$) con configuración *leave-one-out* sobre las muestras de la base de datos, y sobre las que realizamos, a su vez, dos tipos de clasificaciones. La primera basada en la presencia o ausencia de enfermedad y la segunda basada en tres poblaciones diferentes según el grado de enfermedad (AD leve y moderada).

La segunda parte de este capítulo se basa en el uso de clasificadores en Matlab. Los modelos empleados han sido las versiones optimizables de los clasificadores *Tree*, *Discrimintant*, *Naive Bayes*, kNN, SVM y *Ensemble*. Estas versiones optimizables automatizan la selección de valores de parámetros internos del modelo o hiperparámetros como, por ejemplo, el número máximo de divisiones para un árbol de decisión o el *box constraint* de una SVM. Para un tipo de modelo determinado, la aplicación prueba diferentes combinaciones de valores de hiperparámetros mediante un esquema de optimización que busca minimizar el error de clasificación del modelo y devuelve un modelo con los hiperparámetros optimizados.

Para llevar a cabo la clasificación, previamente se ha sometido a las muestras a un proceso de selección de características atendiendo al tipo de variable: temporal, de Temperatura Emocional y conjunto de ambas. Este proceso se ha basado en el uso de una función de selección de características (*feature selection*) mediante el análisis de componentes de vecindad para la clasificación (*fscnca).* Como punto de partida para la búsqueda de las mejores características hemos aplicado una máquina de vectores soporte optimizada y validación cruzada basada en 5 *folds*. Obtenido el que hemos denominado el conjunto óptimo de características hemos procedido a realizar la clasificación estableciendo, según el tamaño de nuestra base de datos, de nuevo, validación cruzada basada en 5 *folds*. Todos los modelos se han entrenado tanto para el conjunto óptimo como para el set total de diez medidas. Asimismo, como ya se hizo en el caso de la clasificación realizada con Stata, se ha realizado idéntico estudio tanto para la clasificación basada en presencia o ausencia de enfermedad como para la basada en grados.

Debido a que se han realizado diferentes análisis y clasificaciones, a modo de resumen, en la Tabla 3-1 se muestra la relación de apartados donde pueden consultarse los resultados obtenidos para cada uno de ellos.

3.6 Discusión y conclusiones

Por último, se analiza de forma crítica el trabajo realizado y los valores obtenidos. A partir de lo expuesto en esta memoria, la discusión comienza valorando el estado del arte y bases de datos expuestas, la situación actual, tendencia y posibles deficiencias.

Respecto a la parte experimental, se relacionan los resultados empíricos de la investigación con aspectos del marco teórico, al tiempo que se presenta una interpretación pormenorizada y análisis crítico para cada uno de ellos.

Por último, se incluyen las conclusiones alcanzadas y líneas de trabajo futuras. Se recogen las ideas y principales hallazgos identificados en la tesis y se hace una recopilación de los nuevos interrogantes planteados junto con posibles líneas de mejora.

Tabla 3-1 Relación de apartados donde encontrar los resultados obtenidos del análisis estadístico, análisis discriminante y clasificadores realizados o entrenados para la comparación de muestras del entrevistadores automático y humano.

					Temporales		Temperatura Emocional (TE)		Combinación Temporales + TE	
					Enfermedad	Grado	Enfermedad	Grado	Enfermedad	Grado
Univariante	**Estadístico**	**Stata**	Estadística descriptiva		7.1.1		7.2.1		-	-
			Paramétrico		7.1.2		7.2.2		-	-
			No paramétrico	Wilcoxon	7.1.3		7.2.3		-	-
				K. Wallis	7.1.3		7.2.3		-	-
				Mediana	7.1.3		7.2.3		-	-
Multivariante			MANOVA		7.1.4		7.2.4		7.3	
	Clasificadores		LDA		8.1.1		8.1.2		8.1.3	
			kNN		8.1.1		8.1.2		8.1.3	
			Logístico		8.1.1		8.1.2		8.1.3	
		Matlab	Tree		-	-	-	-	8.2.2	
			Naive Bayes		-	-	-	-	8.2.2	
			Discriminant		-	-	-	-	8.2.2	
			kNN		-	-	-	-	8.2.2	
			SVM		-	-	-	-	8.2.2	
			Ensemble		-	-	-	-	8.2.2	

Tenemos que, por ejemplo, el análisis de la varianza MANOVA se ha llevado a cabo con el *software* Stata. Se trata de un análisis estadístico multivariante y se ha aplicado sobre cada conjunto completo de características: conjunto de medidas temporales (puede consultarse en el apartado 7.1.4), conjunto de medidas de Temperatura Emocional (puede consultarse en el apartado 7.2.4) y la combinación de ambos conjuntos de medidas (puede consultarse en el apartado 7.3).

Capítulo 4 Base de datos

4.1 Bases de datos en el procesado automático de voz aplicado a la enfermedad de Alzheimer

La disponibilidad de datos que pueden ser objeto de análisis constituye una parte fundamental en cualquier estudio. Para el caso de la detección precoz de la AD a partir de voz, es imprescindible disponer de bases de datos con un número de muestras suficientemente amplio y representativo de los objetos de estudio, en este caso, de pacientes de Alzheimer de diferentes grados (leve, moderado y severo) y sujetos HC.

Hasta hoy, en el estudio de la AD se han empleado las más diversas técnicas atendiendo a diferentes funciones cerebrales y conductuales. Es importante destacar, asimismo, que una de las principales limitaciones que encuentran este tipo de estudios ha sido la escasez y, a su vez, heterogeneidad de muestras disponibles para entrenar modelos que permitan el control evolutivo de la AD. Y es que, la mayor parte de las bases de datos localizadas, carecen de la cantidad necesaria de datos para realizar análisis verdaderamente consistentes, con el inconveniente añadido de haberse llevado a cabo siguiendo diferentes pautas y criterios. Las bases de datos localizadas en esta tesis, en su mayoría, se centran en grabaciones de habla espontánea [111], [137], [199], aunque existen otras que invitan al sujeto a realizar tareas como repetición o lectura [200]. La grabación de las muestras, por norma general, se lleva a cabo de manera puntual sobre un número de sujetos de control y pacientes AD clasificados atendiendo a diferentes grados de severidad: leve, moderado y severo [201]–[204]. Aunque menos, algunos han llevado a cabo estudios longitudinales sobre los sujetos [205], [206] con el fin de encontrar una relación entre el curso de la AD y el deterioro del lenguaje.

A modo de revisión del estado del arte, las bases de datos localizadas que incluyen grabaciones de voz de pacientes de AD se relacionan en la Tabla 4-1. En ella se incluye información destinada a orientar al lector y los estudios relacionados sólo se etiquetan con fines informativos en lugar de ser una clasificación en sí misma. Se ofrece información sobre el tipo de entrevistador utilizado en ellas: humano/automático, y

sobre el tipo de estudio en el tiempo: longitudinal/transversal. Otro tipo de información también incluida es el idioma empleado por los sujetos, la clasificación de los participantes (HC, MCI, AD) y las tareas lingüísticas llevadas a cabo.

Tabla 4-1 Bases de datos de grabaciones utilizadas para análisis lingüístico de AD localizadas y tipos de estudio atendiendo a la distribución en el tiempo de la toma de medidas y al tipo de entrevistador utilizado [36], [207].

Base de datos	**Entrevistador**	**Long/Transv**	**Idioma**	**Tarea**	**HC**	**MCI**	**AD**	**Referencias**
					M/F	**M/F**	**M/F**	
SAIOTEK	Humano	Transversal	Varios	SS	5	-	3	[88], [125]
AZTIAHO				SS	50	-	20	[90], [168], [166], [89], [173], [189]
PGA-OREKA				cVF	26/36	-	17/21	[186]
MINI-PGA				SS	12		6	[115]
-	Humano	Transversal	Griego	SS	30	-	30	[105]
Dem@care	Humano	Transversal	Griego	Varios	4/15	12/31	3/24	[95]
		Transversal	Francés	Varios	6/9	11/12	13/13	[103], [109], [110]
-	Humano	Transversal	Francés	Lectura	14	14	14	[104]
-	Humano	Transversal	Francés	SS	5	-	5	[208]
*TRANSC	Humano	Longitudinal	-	SS	184*	-	214*	[209]
ClinicalTrials.gov	Humano	Longitudinal	Inglés	SS	27	14	-	[139]
-	Humano	Transversal	Inglés	SS	46	-	26	[130]
WRAP	Humano	Longitudinal	Inglés	SS	200	-	64	[114]
Pitt (DB)	Humano	Longitudinal	Inglés	SS	74	19	169	[111], [113], [137], [135], [144], [141], [145], [199], [206]
Kempler (DB)	Humano	Transversal	Inglés	SS	-	-	6	
Lu (DB)	Humano	Transversal	Mandarino	SS	-	-	52	
Lu (DB)	Humano	Transversal	Taiwanés	SS	-	-	16	
PerLA (DB)	Humano	Transversal	Español	SS	-	-	21	
ACADIE	Humano	Longitudinal	Inglés	SS	-	-	95	[133]
AMI	Humano	Transversal	Inglés	SS	20	-	20	[145]
CCC	Humano	Transversal	Inglés	SS	10	-	55	[141], [210]
ILSE	Humano	Longitudinal	Alemán	SS	80	13	5	[129]
CREA-IMSERSO	Humano	Transversal	Español	Lectura	-	-	21	[94]
-	Humano	Transversal	Español	Lectura	82	-	45	[106]
-	Humano	Transversal	Español	Varios	29	-	34	[131], [211]
Cinderella	Humano	Transversal	Portugués	SS	20	20	20	[148]
OPLON	Humano	Transversal	Italiano	Varios	48	48		[101]
-	Humano	Transversal	Iraní	SS	15/15	-	16/14	[98]
-	Automático	Transversal	Japonés	SS	7/3	-	9/1	[99], [100]
-	Humano	Transversal	Japonés	SS	73/200			[107]
BELBI	Humano	Transversal	Serbio	SS	-	-	12	[212]
BEA	Humano	Transversal	Húngaro	SS	13/23	16/32	-	[108], [132], [213], [214]
-	Humano	Transversal	Turco	SS	31/20	-	18/10	[138]
-	Humano	Transversal	Francés	SS	29	-	29	[215]
-	Humano	Transversal	Persa	SS	0/6	-	0/6	[216], [217]
Otros estudios cuyas bases de datos no hemos podido definir: [97], [102], [146], [147], [165].								

HC: sujeto de control, M: masculino, F: femenino, MCI: *Mild Cognitive Impairment*, DB: *DementiaBan*k, PC: *Pitt Corpus*, CCC: *Carolina Conversation Collection*, ILSE: *Interdisciplinary Longitudinal Study of Adult Development and* Aging, ACADIE: *Atlantic Canada Alzheimer's Disease Investigation of Expectations*, WRAP: *Wisconsin Registry for Alzheimer's Prevention*, BEA: *Hungarian Spoken Language Database.* *Transcripciones, no participantes.

A pesar de que los estudios localizados son muy heterogéneos, tanto en términos de las poblaciones que se consideran como de las tareas que realizan, esta relación de bases de datos específicamente centrada en la AD no ha tenido en cuenta exámenes más profundos, por ejemplo, en cuanto a una clasificación detallada de los subtipos de MCI. En este sentido, como primera aproximación, no se analizan en esta tesis los diferentes subtipos que podrían tener impedimentos específicos en los grupos HC, MCI o AD.

Con respecto a las tareas realizadas por los sujetos grabados y características obtenidas de estas tareas, en este trabajo no se profundiza en el análisis de los diferentes significados que una misma característica obtenida de dos tareas orales puede tener (un ejemplo de ello puede ser el significado de una pausa hecha en una tarea de habla espontánea, la cual no tiene necesariamente el mismo significado que una pausa en una tarea de descripción de imagen).

Para la clasificación realizada en esta tesis se ha considerado que las tareas de habla espontánea (SS) implican todas las tareas en las que se le pregunta al sujeto sobre un tema específico y responde libremente durante un cierto período de tiempo relativamente largo. Esto puede incluir desde una amplia gama de entrevistas a descripciones de imágenes. En este sentido es interesante recalcar que, aunque podamos suponer que las pausas realizadas durante las tareas de fluidez verbal pueden ser indicativas de problemas de memoria semántica, esto no es necesariamente cierto para las pausas realizadas durante una tarea de descripción de imágenes. Es necesario aclarar que esta revisión no desarrolla o analiza este punto. Por su parte, tareas de lectura, repetición, conteo o denominación de animales (cVF) son otras pruebas que aparecen en algunos estudios, aunque menos, y entendemos que no tienen relación con el concepto de habla espontánea, tal y como aquí lo definimos. Hemos empleado el término *varios* cuando se han incluido en el estudio tanto tareas SS como otras.

Una de las principales bases de datos que aparecen en la Tabla 4-1 es *Atlantic Canada Alzheimer's Disease Investigation of Expectations* (ACADIE). Este repositorio incluye 2 entrevistas por paciente tomadas con 12 semanas de diferencia para 79 sujetos (AD y HC): leve (50), moderado (53), severo (20) y controles (35). Llevado a cabo en inglés en el Atlántico de Canadá [203], [204], en una de las regiones más envejecidas de este país.

Por su parte, el *Interdisciplinary Longitudinal Study of Adult Development and Aging* (ILSE) es un estudio desarrollado en Alemania cuya base de datos contiene más de 8.000 horas de entrevistas biográficas y diagnósticos cognitivos (AD, MCI, AACD, HC) de más de 1.000 participantes en el transcurso de 20 años. Sólo se han transcrito 380 horas [205].

TALKBANK constituye un conjunto de bases de datos públicamente disponible para cada subcampo de la comunicación. Centrados en aspectos clínicos encontramos: DementiaBank, RHDBank, TBIBank, AphasiaBank, ASDBank, FluencyBank y DementiaBank [206], que incluye Pitt Corpus. Esta última es una base de datos producida por sujetos diagnosticados con demencia y sujetos HC. Contiene entrevistas transcritas y utiliza la tarea de descripción de imagen "Robo de galletas" [76] del Boston Diagnostic Aphasia Examination [218].

Otra base de datos que hemos localizado, AMI, contiene entrevistas biográficas de pacientes con AD y HC. Cada entrevista consta de 4 historias sobre acontecimientos de un período de la vida del sujeto. Contiene datos multimodales y consta de 100 horas de grabaciones [202].

El proyecto europeo Dem@Care [219], por su parte, abarca diferentes aspectos del paciente entre los que se incluye el análisis del habla. Contiene diferentes bases de datos como las descritas en la Tabla 4-1.

También hemos encontrado corpus en los que se recogen diferentes idiomas. Propiedad de la Universidad Médica de Carolina del Sur, Colección de Conversaciones de Carolina (CCC), es una colección digital de grabaciones de audio y video protegida por contraseña y transcrita a partir del habla espontánea de los sujetos [210]. Tiene dos cohortes: 125 hablantes mayores multiétnicos, la mayoría de Carolina del Norte sin discapacidad con cualquiera de hasta 12 afecciones crónicas descritas en el estudio y contiene 70,45 horas de grabaciones [220].

Desde España, el Proyecto Gipuzkoa-Alzheimer (PGA), incluye varias bases de datos que, a su vez, han definido varios subconjuntos:

- AZTIAHO: habla espontánea multicultural y multilingüe en inglés, francés, español, catalán, vasco, chino, árabe y portugués. Comprende también las grabaciones de video HC: 50 (12 horas), AD: 20 (8 horas) [201]. El subconjunto AZTIAHORE [115] consta para el grupo HC: 20 (9 horas), AD: 20 (60 minutos).
- PGA-OREKA: CVF (*Animal naming*, AN). HC: 187, MCI: 38 forman un subconjunto de PGA [115], [201].
- MINI-PGA: descripción de imágenes. HC: 12, AD: 6 [115].

Otra de las bases de datos localizadas es Wisconsin Registry for Alzheimer's Prevention (WRAP). Se trata de un estudio longitudinal con personas con antecedentes parentales de AD. Desde finales de 2001 este estudio ha inscrito a 1561 personas con una edad media inicial de 54 años. Los participantes realizan una segunda visita 4 años después del inicio del estudio, y las visitas posteriores se realizan cada 2 años. El 81% de los participantes permanece activo en el estudio tras 9 años de seguimiento [149].

Para concluir el repaso, Hungarian Spoken Language Database (BEA) es una base de datos multipropósito en húngaro que abarca tareas de habla espontánea, repetición de oraciones y lectura [171]. Consta de 260 horas producidas por 280 hablantes de Budapest (edades entre 20 y 90, 168 mujeres y 112 hombres), proporcionando también anotaciones para distintas investigaciones y aplicaciones prácticas, entre ellas MCI.

A partir de la revisión realizada hemos podido comprobar que un denominador común en la recogida de estas muestras es recurrir a una persona que dirija la entrevista, normalmente incitando al sujeto a evocar recuerdos pasados o a responder preguntas [137], [199], leer textos [221] o describir imágenes [222]. Sin embargo, algunos ya han incluido en sus trabajos avatares informáticos [107], [108] capaces de guiar conversaciones y que, en principio, podrían dotar a las entrevistas de resultados más

objetivos y menos dependientes de factores humanos como puede ser la habilidad o empatía del entrevistador para llevar a cabo una entrevista. Sea como sea, sí que parece relativamente necesario que, ante tanta heterogeneidad en el proceso de recogida de muestras, se pueda llegar a un método más general y objetivo de extracción de habla para pacientes AD.

Varios esfuerzos en investigación han identificado ventajas relevantes para la comunicación asistida por ordenador en comparación con la interacción humana. Partiendo de la base de que el sistema está capacitado con habilidades comunicativas que facilitan dicha comunicación, una de estas ventajas es que el entrevistado siente más anonimato durante el proceso. Para disponer de estas habilidades comunicativas, los sistemas utilizan la visión y el análisis prosódico para implementar comportamientos de escucha activa así como sonrisas, movimientos de cabeza y mímica postural. Además, utilizan estrategias no verbales y muestras de empatía para crear comodidad en el sujeto entrevistado, lo que produce que la conversación sea más fluida. Esto ayuda a generar sentimientos de simpatía y confianza en el entrevistado [223]. Es necesario, pues, que el entrevistador automático sea capaz de percibir el comportamiento humano para procesarlo, comprenderlo y efectuar la consecuente reacción al mismo. También es necesario que estas habilidades no funcionen de forma aislada y que permitan la integración en otros sistemas para que, conforme son utilizados, aprendan de la interacción humana.

A día de hoy, hay sistemas que trabajan para reducir las barreras de los entrevistadores y sistemas virtuales que interactúan con humanos. Implementando sistemas de reconocimiento de voz automáticos [224], el reconocimiento de expresiones faciales [225] y la generación de lenguaje natural [226]. Los esfuerzos de estandarización han intentado consolidar todos los sistemas. Por ejemplo SAIBA [227] es un *framework* para generar conversaciones empáticas. Por otro lado, pueden encontrarse herramientas como LiteBody y DTAsk que crean relaciones socioemocionales con los usuarios mediante comportamientos sociales verbales y no verbales [228]. Por su parte, GRETA [229] es un agente conversacional que cumple con SAIBA, que se enfoca menos en la interacción del lenguaje natural y más en la generación y realización de comportamientos no verbales afectivos. En particular, emplea un modelo de generación facial compleja. El proyecto SEMAINE [230] tiene como objetivo integrar varias tecnologías de investigación, incluidas algunas de las anteriores, en la creación de un oyente virtual. El énfasis está en la percepción y la canalización inversa en lugar de representaciones profundas del diálogo. Incluso hay herramientas que permiten la creación de humanos virtuales que interactúan con las personas en una conversación [231]. Un ejemplo de entrevistador automático que se basa en un humano virtual y que emplea múltiples recursos para que la comunicación sea lo más humana posible es SimSensei Kiosk [232].

A pesar de estos avances en el campo, pocos son los trabajos que se han centrado en aplicar este tipo estrategias, al menos, aplicadas a la AD. Aunque hay constancia de algunos estudios como SimpleC, la solución desarrollada por IBM Watson [82] que sustituye las variables meramente humanas, éstos siguen siendo escasos. A día de hoy, no se ha demostrado hasta qué punto para un mismo sujeto, los

parámetros obtenidos a partir de un entrevistador humano o automático pueden variar y cómo sería posible encajar verdaderamente esta tecnología con la realidad actual.

4.2 *Software* PROGNOSIS

Desarrollado por la Universidad de Las Palmas de Gran Canaria, el *software* Prognosis es un sistema conversacional automático cuyo principal objetivo es realizar grabaciones basadas en el concepto "entrevistador automático". Esta herramienta guía durante el proceso de entrevista a los participantes utilizando una serie de instrucciones y vídeos autoguiados. Concretamente, muestra tres vídeos y, automáticamente, a medida que el participante responde, graba sus comentarios de manera automática. El programa se divide en tres fases principales (ver Figura 4-1).

Para obtener las muestras de habla inducida e incitar a los sujetos a hablar, se ha utilizado como estímulo una serie de vídeos teniendo en cuenta que el denominador común de todos los participantes es la edad y que la gran mayoría supera los 65 años. En el Portal Memoria Digital de Canarias de la Biblioteca Universitaria de la ULPGC [233] hay un amplio repositorio de reportajes y noticias antiguas, entre otras, que hemos utilizado para provocar reminiscencias en los sujetos.

Actualmente, existen numerosos estudios que demuestran la efectividad del uso de herramientas multimedia, como los vídeos, para hacer que los sujetos que padecen demencia o AD se comuniquen o expresen con mayor facilidad [234]–[237], y también son especialmente interesantes cuando se personalizan o provocan reminiscencias de momentos vividos. En base a esto, hemos seleccionado expresamente del repositorio aquellos vídeos que reflejan tiempos pasados y que, por lo tanto, consideramos que podrían despertar recuerdos en los participantes.

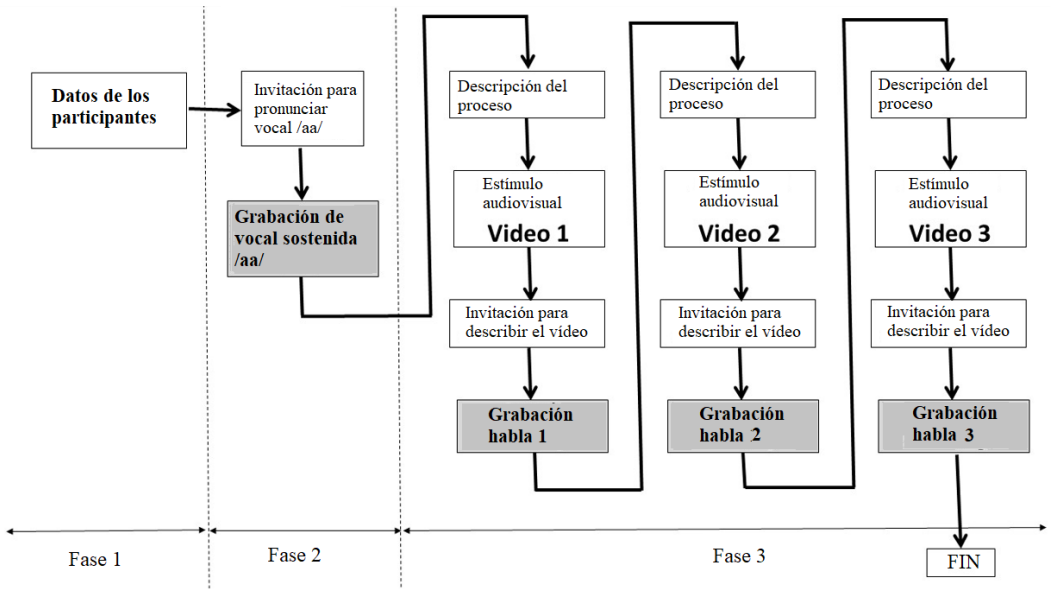


Figura 4-1 Diagrama de flujo *software* PROGNOSIS [207].

La primera fase del programa corresponde al registro de los datos del participante. Las entrevistas son anónimas, pero se recogen algunos datos de los participantes como el grado de AD (control, leve o moderado), sexo y edad.

Iniciada la aplicación, lo primero que se muestra es una ventana general (Figura 4-2) con las instituciones que participan en el proyecto y la posibilidad de elegir idioma. A continuación, en una segunda ventana (Figura 4-3), se pide detallar los datos personales del sujeto a estudiar, a fin de establecer unas etiquetas unívocas cuya estructura coincida con lo explicado en el apartado 3.1.

Figura 4-2 Primera ventana del *software* Prognosis

Figura 4-3 Ventana para introducir datos personales del sujeto.

Realizados estos ajustes, el *software* muestra un vídeo de uno de los investigadores del proyecto explicando cómo va a ser la grabación y de qué manera han de actuar los sujetos (Figura 4-4). Para empezar, el sujeto deberá pronunciar de forma sostenida la vocal "a". Posteriormente, se reproducirán tres vídeos elegidos aleatoriamente a partir de un repositorio descrito [233] y se grabarán las impresiones y comentarios de los entrevistados. Durante el tiempo en que se esté grabando a los sujetos, se presentará en pantalla una imagen que indica el tiempo grabado (Figura 4-5).

Figura 4-4 Fotograma del video explicativo del proceso.

Figura 4-5 Ventana que indica el tiempo grabado.

Finalizadas las grabaciones, el investigador que guía durante la entrevista da las gracias al sujeto por participar en el estudio, y se vuelve a reiniciar todo el proceso a partir de la ventana de datos personales.

4.3 Cross-Sectional Alzheimer Prognosis R2019

Dada la importancia que tienen las bases de datos en cualquier tipo de estudio, uno de los objetivos principales de esta tesis es disponer de una base de datos suficientemente grande que nos permita realizar diferentes estudios estadísticos y obtener resultados y conclusiones en base a ella.

A partir de lo expuesto en el apartado anterior, en este trabajo se ha querido profundizar en las posibles diferencias que pueden existir en grabaciones de voz tomadas mediante dos metodologías diferentes, ¿existen realmente diferencias a la hora de discriminar AD según el entrevistador utilizado en las grabaciones? Para responder a esta cuestión, entre otras, la base de datos creada está formada por dos tipos de muestras: las obtenidas a partir de un entrevistador automático y las obtenidas a partir de un entrevistador humano. En adelante, ambos tipos de muestras recibirán el nombre de muestras de habla inducida y muestras de habla espontánea, respectivamente.

Esta base de datos a disposición de la Universidad de Las Palmas de Gran Canaria se ha llamado Cross-Sectional Alzheimer Prognosis R2019 (CSAP-R19). De manera general, la base de datos CSAP-R19 está compuesta principalmente por personas mayores de 65 años. Consta de 41 pacientes con AD (31% leve y 15% moderado) y de 46 sujetos HC. De todos ellos, el 64% son mujeres y el 36% son hombres. En total han sido grabados 87 participantes y se han recogido un total de 322 muestras/grabaciones. Aunque no en todos los casos, de cada participante se han obtenido generalmente tres muestras de habla inducida, mediante el *software* Prognosis, y una muestra de habla espontánea, obtenida mediante un entrevistador humano. En la Tabla 4-2, se muestra el balance de sujetos sometidos a estudio en base a la presencia o no de enfermedad, grado, sexo y edad.

Tabla 4-2 Balance por presencia o ausencia de enfermedad, grado, sexo y edad de los participantes.

Balance AD		**Balance por sexo**		**Balance por edad**	
Sujetos AD	41	Mujeres	56	<65	7
AD Leve (AD1)	26	HC	27	65 < x < 80	52
AD Moderado (AD2)	15	AD leve (AD1)	19	>80	11
		AD moderado (AD2)	10	Edad desconocida	17
Sujetos HC	46	Hombres	31		
		HC	22		
		AD leve (AD1)	7		
		AD moderado (AD2)	2		
Total participantes	87				

Asimismo, en la Figura 4-6 y Figura 4-7 se representan de manera visual dichos balances. En el caso de la Figura 4-6, puede observarse tanto el balance por grado de enfermedad de los participantes como por sexo (siendo gris: hombres, azul: mujeres).

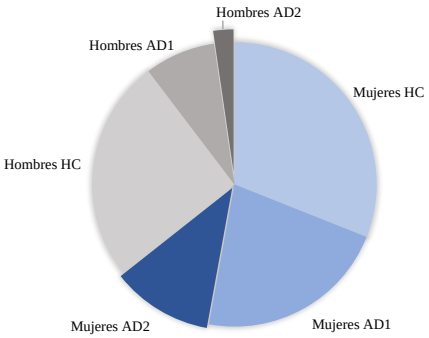


Figura 4-6 Balance de participantes por sexo y grado de enfermedad.

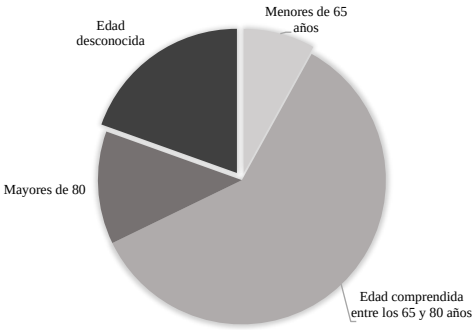


Figura 4-7 Balance de participantes por edad.

Capítulo 5 Extracción de características: medidas temporales

5.1 Introducción

Para llevar a cabo el proceso de extracción de características, el primer paso a realizar es el preprocesado de las muestras cuyo objetivo es preparar las grabaciones para su posterior tratamiento. Para ello, se ha hecho uso del *software* Audacity® y se han escuchado una a una todas las grabaciones. En aquellos casos donde se han identificado sonidos indeseados, por ejemplo, el producido por alguna fuente externa, han sido atenuados hasta quedar enmascarados por el ruido ambiente.

Una vez adaptadas las muestras, se ha hecho uso de un *software* detector de actividad de voz (VAD) [126] basado en un algoritmo implementado en Matlab® por la División de Procesado Digital de Señales del Instituto para el Desarrollo Tecnológico y la Innovación en Comunicaciones (IDeTIC). La finalidad del mismo es analizar toda la grabación y seleccionar los fragmentos de audio (habla) y diferenciarlos del resto de la grabación. Esto permite cuantificar el tiempo de locución que hay en una grabación y diferenciarlo del resto, que puede ser silencio o ruido de fondo. El resultado de este proceso es un conjunto de vectores donde se señala cada una de las partes de audio de la grabación. A partir de los índices iniciales y finales obtenidos por el VAD de cada muestra, se extraen los parámetros de interés.

El proceso discriminatorio del habla frente al silencio se realiza concretamente por tramas. Se evalúa cada trama para quedarse con los segmentos de audio mientras que se descartan aquéllos de silencio. Finalmente, los resultados de cada trama se almacenan en un vector de índices indicando el inicio y fin de cada fragmento de audio. Para saber la duración de la trama sonora, basta con hacer una resta de cada índice final e inicial. A continuación, cada resultado se debe pasar al dominio del tiempo, dividiendo entre la frecuencia de muestreo. Este proceso se ha repetido también para las

tramas de silencio. En la Figura 5-1 se muestra un ejemplo de VAD aplicado a una señal de voz.

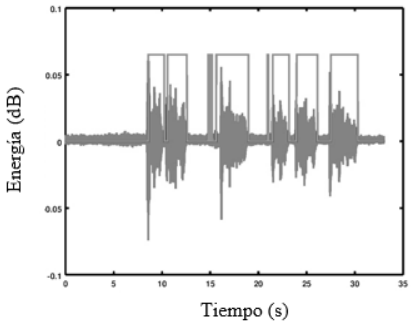


Figura 5-1 Ejemplo *Voice Activity Detector* sobre señal de voz.

Es importante tener en cuenta que la señal de voz no es estacionaria, pero analizada en ventanas cortas (duración entre 10 - 40 ms) sí se puede considerar estacionaria. En esta tesis, la señal ha sido procesada en ventanas cortas, simplificando el análisis debido a que se supone que sus parámetros acústicos son constantes.

Por su parte, con el objetivo de que las medidas fueran lo más precisas posible, se ha añadido un intervalo de guarda de 50 ms al inicio y al final de cada secuencia. Con esta medida se pretende que sonidos tales como chasquidos, de muy corta duración, queden descartados.

5.2 Estadística descriptiva aplicada a tiempos de habla

Se pretende estudiar los tiempos de habla de los locutores grabados como elemento para discriminar entre sujetos sanos y patológicos AD. Para ello se propone calcular el tiempo de habla de cada muestra y caracterizarlo por medio de medidas como el tiempo medio de habla ($\overline{t_S}$), varianza del tiempo de habla ($\sigma_{t_S}^2$), *skewness* del tiempo de habla ($\tilde{\mu}_{t_{S_3}}$), curtosis del tiempo de habla ($Kurt_{t_s}$) y una medida relacional como el índice de tiempo de habla (Ind_{t_s}). Cada una de estas medidas se definen a continuación:

- Tiempo medio de habla ($\overline{t_S}$): describe el tiempo medio de habla de los diferentes fragmentos de sonido que pueden identificarse en una muestra. Se estima a partir de la media aritmética de la duración de todos los fragmentos de sonido en una sola grabación [238].

$$\overline{t_S} = \frac{\sum_{i=1}^{N} t_{S_i}}{N} \qquad \text{(Ec. 1)}$$

  Donde t_{S_i} es la duración de cada fragmento sonoro $(S_1, S_2, \dots, S_N)$ es lo que se divide en cada grabación de voz $\{S_i\}$.

- Varianza del tiempo de habla ($\sigma_{t_S}^2$): describe la variación de los diferentes fragmentos de sonido en una grabación. Se estima [239] utilizando el siguiente estimador de la varianza:

$$\sigma_{t_S}^2 = \frac{\sum_{i=1}^{N}(t_{S_i} - \overline{t_S})^2}{N-1} \qquad \text{(Ec. 2)}$$

- *Skewness* del tiempo de habla ($\tilde{\mu}_{t_{S_3}}$): Esta medida permite caracterizar el comportamiento de la función de distribución de probabilidad de la duración de los fragmentos sonoros. Esta medida cuantifica [240] la falta de simetría respecto a la duración media de los fragmentos de voz. De esta forma, cuando las muestras estudiadas siguen una distribución normal, el valor de $\tilde{\mu}_{t_{S_3}}$ es cero. Valores positivos o negativos de $\tilde{\mu}_{t_{S_3}}$ indica datos sesgados a la derecha de la curva de distribución o a la izquierda, respectivamente. La asimetría del tiempo de habla se calcula utilizando el siguiente estimador:

$$\tilde{\mu}_{t_{S_3}} = \frac{\sum_{i=1}^{N}(t_{S_i} - \overline{t_S})^3}{N \cdot \left(\sqrt{\sigma_{t_S}^2}\right)^3} \qquad \text{(Ec. 3)}$$

  Donde t_{S_i} es la duración de cada fragmento de sonido, $\overline{t_S}$ es el tiempo de habla promedio, $\sigma_{t_S}^2$ es la varianza del tiempo de habla y N es el número de fragmentos de sonido de la muestra.
- Curtosis del tiempo de habla ($Kurt_{t_s}$): Esta es otra medida que permite caracterizar el comportamiento de la función de distribución de probabilidad de la duración de los fragmentos sonoros. Indica la cantidad de fragmentos de voz cuya duración es cercana a la duración media $\bar{h}$. Cuanto mayor sea $Kurt_{t_s}$ más pronunciada será su curva de distribución. $Kurt_{t_s}$ se calcula [240] utilizando la siguiente fórmula:

$$Kurt_{t_s} = \frac{\sum_{i=1}^{N}(t_{S_i} - \overline{t_S})^4}{N \cdot (\sqrt{\sigma_{t_S}^2})^4} - 3 \qquad \text{(Ec. 4)}$$

- Índice de tiempo de habla (Ind_{t_s}): describe la relación entre el tiempo total que el sujeto está hablando y la duración total de la grabación. Se calcula como la división entre el tiempo total de las secuencias de voz por el tiempo total de grabación de la muestra.

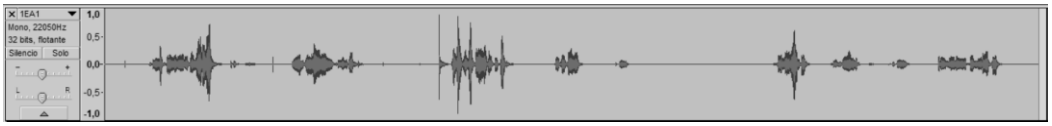
$$Ind_{t_s} = \frac{\sum_{i=1}^{N} t_{S_i}}{T_{TOTAL}} \qquad \text{(Ec. 5)}$$

  Donde t_{S_i} es la duracion de cada fragmento sonoro $(S_1, S_2, \dots, S_N)$ en el que se divide la grabación y T_{TOTAL} es la duración de la grabación completa.

Una vez se han obtenido los parámetros anteriores para cada muestra de audio, se almacenan en un documento *.txt*, tal y como se muestra en la Figura 5-2.

Para conocer exactamente qué variables de las anteriores son discriminantes AD y cuáles no y, además, conocerlo para cada tipo de entrevistador estudiado en esta tesis

(humano o automático), se han realizado diferentes análisis estadísticos que pueden consultarse en el Capítulo 7 y Capítulo 8.



1.5011046e+00 1.3952003e-01 2.9704255e+00 3.3317298e-02 1.2552239e+01 1.0754215e+00

Figura 5-2 Ejemplo de muestra de audio, a partir de un archivo WAV y los valores de las variables descritas anteriormente en formato texto (*.txt*).

Capítulo 6 Extracción de características: medidas de Temperatura Emocional

6.1 Introducción

Tal y como se expuso en el apartado 1.1.2, la capacidad de respuesta emocional de los pacientes de Alzheimer se ve afectada progresivamente produciendo cambios sociales y de comportamiento. Dado que este síntoma se refleja, a su vez, en el uso lenguaje y la capacidad comunicativa del sujeto, pretendemos caracterizar este fenómeno a partir de la Temperatura Emocional.

La Temperatura Emocional es una característica del habla que se obtiene a partir de la extracción de dos características prosódicas y cuatro paralingüísticas, las cuales están relacionadas con el tono y la energía espectral, respectivamente. A continuación, se describe cómo se ha llevado a cabo el cálculo de cada una de ellas.

6.1.1 Rasgos prosódicos

Dentro de los rasgos sonoros del habla, la frecuencia fundamental de las muestras se considera el principal indicador prosódico y, más concretamente, la entonación dada por el contorno de pitch.

De manera general se puede decir que el contorno de pitch, por ejemplo, de una locución con entonación neutral, comienza con un valor de pitch máximo y, a partir de aquí, empieza a decaer progresivamente de manera que el contorno puede dibujarse como una envolvente global descendente. Cuando las locuciones son de carácter más emocional, esta envolvente global cambia, circunstancia que resulta más significativa para diferentes niveles de activación de las emociones.

Para el cálculo de la TE hemos utilizado como parámetros prosódicos dos coeficientes de regresión lineal a y b que se obtienen de la ecuación que modela el contorno del pitch $p(n)$ a partir de un fragmento de audio $\{w(n)\}$ de *N* muestras [92].

$$a, b \,/\, minimiza\ Error, \quad \text{donde } Error = \sum_{i=1}^{N}(p(i) - (bi + a))^2 \qquad \text{Ec. 6}$$

Los coeficientes a y b se calculan utilizando el método de mínimos cuadrados, donde el coeficiente a representa el pitch, mientras que el coeficiente b está relacionado con la disminución o tendencia del tono. Para su implementación hemos utilizado el algoritmo de estimación de pitch llamado YIN [241]. A modo de ejemplo, en la Figura 6-1 se representan tanto el pitch como el contorno de pitch para una señal de voz.

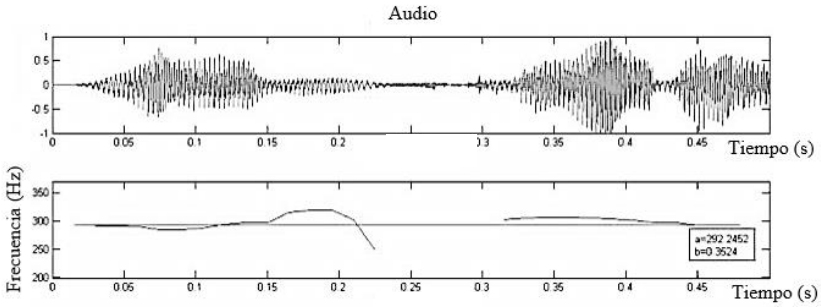


Figura 6-1 Ejemplo de extracción de rasgos prosódicos a partir de la grabación de voz de un sujeto.

6.1.2 Rasgos paralingüísticos

La energía a corto plazo es un indicador prosódico de la tensión léxica del habla. Sin embargo, la acumulación de energía sonora en diferentes bandas de frecuencia, que varía según el modelo de producción del habla, puede usarse también como indicador paralingüístico del estado emocional. En el habla emocional la energía en altas frecuencias aumenta en comparación con un habla carente de emociones.

Para el cálculo de estas características paralingüísticas utilizamos, a partir de la muestra de audio $\{w(n)\}$ [92], cuatro balances de energía espectral de la voz (E_{B_0}, E_{B_1}, E_{B_2} y E_{B_3}). Éstos son cuantificados utilizando cuatro porcentajes de concentración de energía en cuatro bandas de frecuencia B_i (donde $i \in [0, 3]$).

Así, para una frecuencia de muestreo superior a 16 KHz, las bandas de frecuencia se dividen en los rangos B_0 = [0 Hz – 400 Hz], B_1 = [400 Hz – 2 KHz], B_2 = [2 KHz – 5 KHz] y B_3 = [5 KHz – 8 KHz]. El porcentaje de energía en cada banda de frecuencia E_{B_i} se obtiene utilizando la siguiente ecuación:

$$E_{B_i} = \frac{\sum_{f=B_i}|X(f)|^2}{\sum_{f=0}^{8KHz}|X(f)|^2} \qquad \text{(Ec. 7)}$$

Donde $0 \leq i \leq 3$ y $|X(f)|^2$ es un período de la trama temporal de voz $\{w(n)\}$.

En la Figura 6-2 se muestra a modo de ejemplo la representación en frecuencia de una señal de voz en la que pueden apreciarse gráficamente las cuatro bandas de energía definidas.

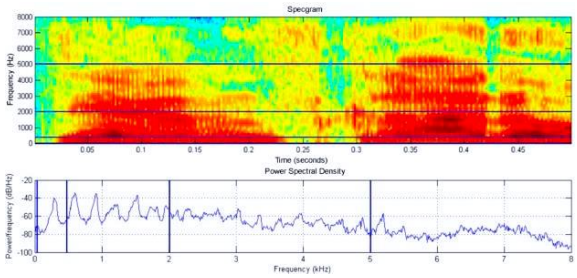


Figura 6-2 Representación de la acumulación de energía en las 4 bandas de frecuencias definidas (E_{B_0}, E_{B_1}, E_{B_2} y E_{B_3}) para una señal de audio.

6.1.3 Cálculo de la Temperatura Emocional

Para llevar a cabo el cálculo de la Temperatura Emocional, en primer lugar se realiza el preprocesado de cada grabación de la base de datos $\{s(n)\}$ mediante un VAD, el cual se encarga de eliminar los silencios de las muestras. De cada señal de voz obtenida se elimina la componente continua (DC) y se realiza una normalización *z-score*. Finalmente, la señal de habla se divide en ventanas (tipo *hamming*) de 0,5 s y *overlapping* o solapamiento del 50%. Como resultado de esta fragmentación temporal se obtienen diferentes segmentos emocionales $\{w(n)\}$ que serán parametrizados mediante las dos características prosódicas y las cuatro paralingüísticas descritas en los apartados 6.1.1 y 6.1.2.

A partir de aquí, el cálculo de la Temperatura Emocional se realiza en dos fases principales. En primer lugar se evalúa cada segmento emocional o vector de parámetros $\{w(i)\}$ con un clasificador tipo *Support Vector Machine* (SVM), concretamente de la librería LIBSVM de Matlab. En esta fase del cálculo de la Temperatura Emocional, cada segmento emocional $\{w(i)\}$ se clasifica mediante la librería LIBSVM en dos clases: activación alta y activación baja. Para ello, el umbral de decisión (*Th1*) se calcula a partir del EER obtenido de los datos de entrenamiento.

En la segunda parte del cálculo, se hace uso del umbral promedio TH_2, dado por el porcentaje de los segmentos anteriores considerados de activación alta. Se calcula a partir del EER obtenido de los datos de validación de la SVM y establece el porcentaje mínimo de segmentos de activación alta que son necesarios para que la señal sea clasificada como tal. Dicho de otra manera, cuando el porcentaje de segmentos emocionales clasificados como de activación alta en la primera etapa, en relación con todos los segmentos emocionales de una señal de habla completa, esté por encima del segundo umbral TH_2, la muestra será clasificada como de activación alta.

A partir de este umbral se realiza una normalización lineal, tal y como se muestra en la Figura 6-3, obteniendo lo que se conoce como la Temperatura Emocional (TE o *ET*, por sus siglas en inglés).

De manera general, una señal de voz se clasifica como de emocional cuando TE $\geq$ 50 y como no emocional cuando $TE \leq 50$. Según el porcentaje de segmentos que se

hayan clasificado como de activación alta, el resultado obtenido para la TE puede tomar el valor de 0 (en el 0%), 50 (en el umbral TH_2) y 100 (en el 100%).

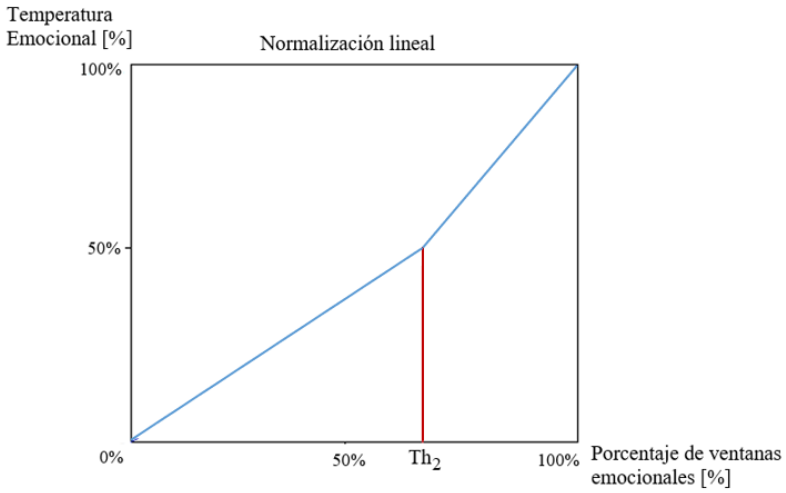


Figura 6-3 Escala de normalización lineal de Temperatura Emocional.

6.2 Estadística descriptiva aplicada a la Temperatura Emocional

Como se ha comentado, para el análisis estadístico que se realizará en los próximos capítulos, se tiene en consideración tanto la Temperatura Emocional discreta (TE_d) como la continua (TE_c).

El cálculo de la Temperatura Emocional discreta (TE_d) ha sido desarrollado en detalle en el apartado 6.1.3 y, a partir ello, lo que obtenemos es un único número natural que ofrece información sobre si el habla de una grabación completa se considera emocional ($TE_d > 50$) o no.

Para el cálculo de los estadísticos descriptivos relacionados con la Temperatura Emocional continua (TE_c), previamente una función divide la grabación en diferentes segmentos de 1 segundo de duración obteniendo, para cada uno de estos segmentos, un valor de Temperatura Emocional. Posteriormente, estos valores se almacenan en un vector. Para caracterizar este vector calculamos sus estadísticos descriptivos; el valor medio [242], varianza [239], curtosis y *skewness* [240] para una grabación determinada. Concretamente, las variables definidas han sido las siguientes:

- Valor medio de la Temperatura Emocional continua ($\overline{te_c}$): es el valor medio de la Temperatura Emocional de los diferentes fragmentos de 1 segundo en los que se divide la muestra. Se calcula a partir de la media aritmética de los valores del vector de Temperatura Emocional continua.

$$\overline{te_c} = \frac{\sum_{i=1}^{N} te_{c_i}}{N} \quad \text{(Ec. 8)}$$

  En donde $\mathrm{te_{c_i}}$ es el valor de TE_c de cada fragmento $(S_1, S_2, \dots, S_N)$ en los que se divide en cada grabación de voz $\{S_i\}$.

- Varianza de la Temperatura Emocional continua $\left(\sigma_{te_c}^2\right)$: describe la variación de los valores almacenados en el vector de Temperatura Emocional continua de los diferentes fragmentos en una grabación. Se calcula [239] utilizando el siguiente estimador de la varianza:

$$\sigma_{te_c}^2 = \frac{\sum_{i=1}^{N}(te_{c_i} - \overline{te_c})^2}{N-1} \quad \text{(Ec. 9)}$$

- *Skewness* de la Temperatura Emocional continua ($\tilde{\mu}_{te_{c_3}}$): *skewness* de los valores del vector de Temperatura Emocional continua. Esta medida permite caracterizar el comportamiento de la función de distribución de probabilidad de los valores del TE_c de los diferentes fragmentos almacenados en el vector. Esta medida cuantifica [240] la falta de simetría respecto al valor medio $\overline{te_c}$ de los fragmentos de voz. De esta forma, cuando las muestras estudiadas siguen una distribución normal, el valor de $\tilde{\mu}_{te_{c_3}}$ es cero. Valores positivos o negativos de $\tilde{\mu}_{te_{c_3}}$indican datos sesgados a la derecha de la curva de distribución o a la izquierda, respectivamente. La asimetría de los valores de Temperatura Emocional se calcula utilizando el siguiente estimador:

$$\tilde{\mu}_{te_{c_3}} = \frac{\sum_{i=1}^{N}(te_{c_i} - \overline{te_c})^3}{N \cdot \left(\sqrt{\sigma_{te_c}^2}\right)^3} \quad \text{(Ec. 10)}$$

  Donde $\mathrm{te_{c_i}}$ es el valor de Temperatura Emocional de cada fragmento de la grabación, $\overline{te_c}$ es el valor promedio, $\sigma_{te_c}^2$ es la varianza de los valores de Temperatura Emocional del vector y N es el número de fragmentos analizados de la muestra.

- Curtosis de la Temperatura Emocional continua ($Kurt_{te_c}$): Esta es otra medida que permite caracterizar el comportamiento de la función de distribución de probabilidad de los valores de TE de los diferentes fragmentos en los que se divide la muestra. Indica la cantidad de fragmentos de voz cuya TE es cercana al valor de Temperatura Emocional medio ($\overline{te_c}$). Cuanto mayor sea $Kurt_{te_c}$ más pronunciada será su curva de distribución. Se calcula [240] utilizando la siguiente fórmula:

$$Kurt_{etc_s} = \frac{\sum_{i=1}^{N}(etc_{S_i} - \overline{etc_S})^4}{N \cdot \left(\sqrt{\sigma_{etc_S}^2}\right)^4} - 3 \quad \text{(Ec. 11)}$$

Una vez se han obtenido las características anteriores para cada grabación, se conforma el set de cinco medidas de Temperatura Emocional descrito al inicio de esta memoria (apartado 3.3) y con el que se pretende caracterizar la carga emocional del habla.

Capítulo 7 Estudio estadístico

Para confirmar la hipótesis inicial en la que planteábamos que es posible discriminar Alzheimer a partir de las grabaciones de voz obtenidas de un sistema conversacional automatizado o entrevistador automático y conocer las capacidades que tiene respecto a su homólogo manual, vamos a caracterizar y analizar estadísticamente las muestras de habla mediante los parámetros temporales y de carga emocional descritos en los Capítulo 5 y Capítulo 6 y la combinación de ambos simultáneamente.

7.1 Medidas temporales

Para saber exactamente qué variables temporales (descritas en el Capítulo 5) son discriminantes AD y cuáles no para cada entrevistador, se ha realizado un análisis estadístico descriptivo, un estudio paramétrico para ver si las muestras reflejan una distribución normal [243] y un estudio no paramétrico basado en tres tipos de pruebas: Test de Wilcoxon, Test Kruskal-Wallis y Test de la Mediana. Por último se ha realizado un análisis multivariante MANOVA para estudiar el comportamiento de las cinco variables como conjunto.

7.1.1 Análisis de estadística descriptiva

Como se ha indicado previamente, en un primer paso se ha realizado un análisis de estadística descriptiva basado específicamente en los valores medio y de desviación estándar sobre cada una de las variables del conjunto de muestras.

En este apartado se hace una clasificación de las muestras atendiendo a cada una de las tres poblaciones bajo estudio: sujetos de control (HC), pacientes de Alzheimer grado leve (AD1) y pacientes de Alzheimer grado moderado (AD2). En el caso de la población denominada AD, se incluyen las muestras relativas a los grados leves y moderados (grupos AD1 y AD2). A fin de facilitar la comparación entre los diferentes

entrevistadores, en la Tabla 7-1 se muestran los valores de las variables obtenidos para cada entrevistador.

Se puede observar que los valores del habla de los sujetos HC frente a los de pacientes AD son, en todos los casos y para los dos entrevistadores, mayores. Las diferencias entre los valores de la población HC y el resto de poblaciones se hace mayor conforme aumenta el grado de enfermedad.

Tabla 7-1 Valores de estadística descriptiva de las medidas temporales de habla para cada población y para cada entrevistador: valor medio (μ) y desviación estándar (σ).

	Poblaciones							
	HC		**AD1**		**AD2**		**AD (AD1 + AD2)**	
Variable/ Entrevist.	**Human.** μ (σ)	**Autom.** μ (σ)	**Human.** μ (σ)	**Autom.** μ (σ)	**Human.** μ (σ)	**Autom.** μ (σ)	**Human.** μ (σ)	**Autom.** μ (σ)
$\bar{t_s}$	1,79 (0,53)	1,8 (0,76)	1,47 (0,29)	1,38 (0,47)	1,19 (0,69)	1,30 (0,42)	1,42 (0,39)	1,36 (0,45)
$\sigma^2_{t_s}$	1,6 (1,32)	1,33 (1,40)	0,94 (0,44)	0,97 (1)	0,88 (0,60)	0,74 (1,04)	0,93 (0,47)	0,9 (1,01)
$\tilde{\mu}_{t_{s_3}}$	0,81 (0,56)	0,49 (0,59)	0,92 (0,47)	0,49 (0,65)	0,80 (0,46)	0,43 (0,63)	0,9 (0,46)	0,47 (0,64)
$Kurt_{t_s}$	3,3 (1,52)	2,3 (1,1)	3,52 (1,34)	2,44 (1,28)	2,77 (1,69)	2,18 (0,92)	3,37 (1,41)	2,35 (1,17)
Ind_{t_s}	0,71 (0,1)	0,7 (0,12)	0,51 (0,08)	0,48 (0,16)	0,38 (0,27)	0,43 (0,21)	0,48 (0,14)	0,47 (0,18)

HC: sujeto de control, AD1: Alzheimer grado leve, AD2: Alzheimer grado moderado.

En base a los datos obtenidos en el proceso de extracción de características, la Tabla 7-2 muestra los diagramas de caja de las muestras analizadas para cada una de las cinco variables que hemos considerado para este estudio.

A partir de ellos es posible ver, para ambos entrevistadores, una mayor capacidad discriminante entre muestras AD y control para la variable $Ind_{t_s.}$ Otras, como $\tilde{\mu}_{t_{s_3}}$ o $Kurt_{t_s}$, resultan ser poco discriminantes sea cual sea el entrevistador utilizado.

7.1.2 Análisis paramétrico

Los resultados obtenidos no siguen una distribución normal por lo que, en este caso, no estaría indicado realizar un análisis paramétrico. A partir de la regresión lineal de todas las variables temporales se observa que la recta de la regresión no se ajusta correctamente a todos los datos. Los resultados del test de normalidad de residuos, en concreto el test de *skewness* y curtosis, nos indica a partir de los valores de *Chi* cuadrado que la regresión no es correcta y que los residuos no seguirían dicha distribución normal.

7.1.3 Análisis no paramétrico

En la Tabla 7-3 se muestran los resultados obtenidos de los tres test no paramétricos llevados a cabo para comparar las variables temporales obtenidas a partir del entrevistador humano y automático.

Los valores en los que la *Prob*|*z*| es mayor de 0,05 (*Prob*|*z*| > *0,05*), casos en los que no hay diferencia al comparar las medidas de las poblaciones, se han representado en color gris. Estos valores avalarían la hipótesis nula y, por tanto, indicarían que una variable no es discriminante AD.

Puede observarse que los peores resultados se obtienen cuando comparamos las dos poblaciones de pacientes de AD: grado leve (AD1) y grado moderado (AD2). En este caso, el número de variables que resultan discriminantes AD se ven reducidas considerablemente.

Tabla 7-2 Diagramas de cajas o *boxplots* para las cinco variables temporales consideradas.

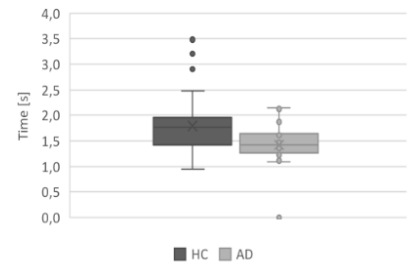
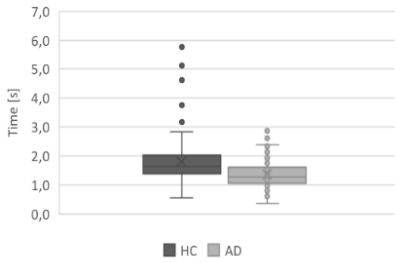
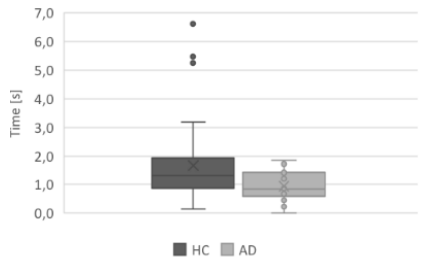
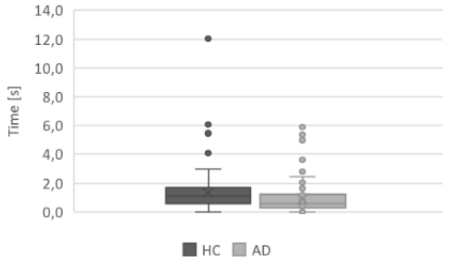
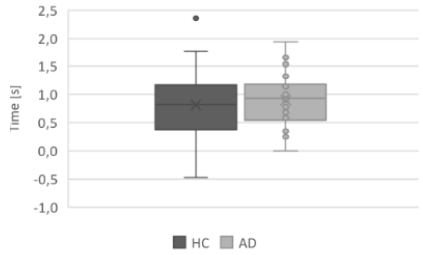
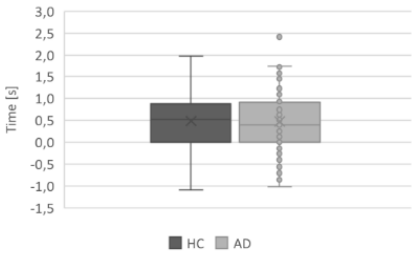
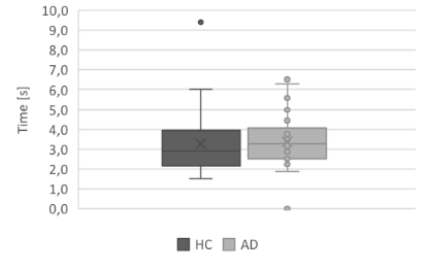
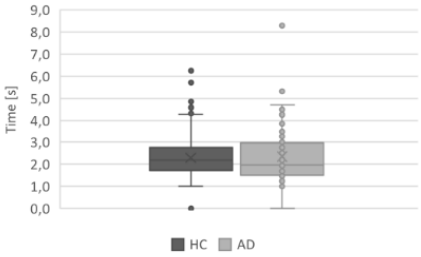
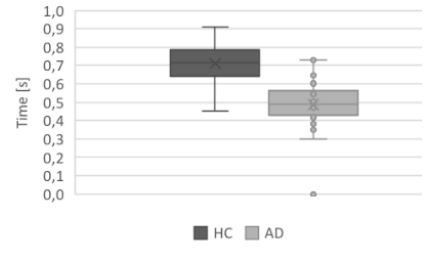
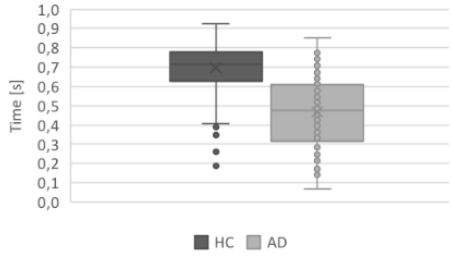
Variable/ Entrevistador	**Humano**	**Automático**
$\bar{t_S}$		
$\sigma^2_{t_S}$		
$\tilde{\mu}_{t_{S_3}}$		
$Kurt_{t_s}$		
Ind_{t_s}		

Tabla 7-3 Resultados análisis univariante no paramétrico: qué variables temporales son o no discriminantes comparando las diferentes poblaciones y entrevistadores.

	Test Wilcoxon		**Test Kruskal-Wallis**		**Test Mediana**	
	Prob|z|		χ^2		**Pearson χ^2**	
Variable/ Entrevist.	**Humano**	**Automático**	**Humano**	**Automático**	**Humano**	**Automático**
HC vs AD						
$\bar{t_S}$	0,002	<1e-10	0,002	<1e-10	0,007	<1e-10
$\sigma^2_{t_S}$	0,005	<1e-10	0,005	<1e-10	0,086	<1e-10
$\tilde{\mu}_{t_{S_3}}$	0,560	0,637	0,560	0,637	0,462	0,525
$Kurt_{t_s}$	0,410	0,515	0,410	0,515	0,462	0,373
Ind_{t_s}	<1e-10	<1e-10	<1e-10	<1e-10	<1e-10	<1e-10
HC vs AD1						
$\bar{t_S}$	0,007	<1e-10	0,007	<1e-10	0,002	<1e-10
$\sigma^2_{t_S}$	0,010	0,006	0,010	0,006	0,137	0,002
$\tilde{\mu}_{t_{S_3}}$	0,510	0,687	0,508	0,687	0,542	0,418
$Kurt_{t_s}$	0,320	0,780	0,317	0,784	0,542	0,418
Ind_{t_s}	<1e-10	<1e-10	<1e-10	<1e-10	<1e-10	<1e-10
HC vs AD2						
$\bar{t_S}$	0,053	<1e-10	0,053	<1e-10	0,066	0,011
$\sigma^2_{t_S}$	0,145	<1e-10	0,145	<1e-10	0,963	<1e-10
$\tilde{\mu}_{t_{S_3}}$	1,000	0,720	1,000	0,729	0,963	0,741
$Kurt_{t_s}$	0,870	0,360	0,874	0,357	0,963	0,284
Ind_{t_s}	0,006	<1e-10	0,006	<1e-10	0,370	<1e-10
AD1 vs AD2						
$\bar{t_S}$	0,047	0,380	0,770	0,385	1,000	0,228
$\sigma^2_{t_S}$	0,283	0,030	0,922	0,039	1,000	0,108
$\tilde{\mu}_{t_{S_3}}$	0,035	0,870	0,974	0,872	1,000	0,688
$Kurt_{t_s}$	0,255	0,520	0,672	0,526	1,000	1,000
Ind_{t_s}	0,871	0,140	0,313	0,145	1,000	0,228

HC: sujeto de control, AD1: Alzheimer grado leve, AD2: Alzheimer grado moderado.

7.1.4 Análisis multivariante MANOVA

Antes de realizar el análisis multivariante MANOVA, es necesario comprobar que las muestras siguen una distribución gaussiana. A tal efecto, hemos realizado una prueba de normalidad basada en asimetría y curtosis.

Una vez verificado que los resultados de la prueba, efectivamente, apoyan la aplicación de un análisis tipo MANOVA, los valores *p* obtenidos del análisis multivariado se muestran en la Tabla 7-4. Los estadísticos *W*, *P*, *L* y *R* listados en la en ella hacen referencia a *Wilks' lambda*, *Pillai's trace*, *Lawley-Hotelling trace* y *Roy's largest root*, respectivamente.

Al igual que en el análisis anterior, en gris, se muestran aquellos estadísticos que indicarían que el conjunto de variables no discrimina AD. Como se puede observar, los cuatro estadísticos indican que las variables dependientes bajo estudio resultarían discriminantes en todos los casos excepto cuando la variable de agrupamiento se basa en el grado de la enfermedad y sólo se consideran las poblaciones leves y moderadas (AD1 y AD2).

Tabla 7-4 Resultados del análisis multivariante MANOVA para los estadísticos *Wilks' lambda*, *Lawley-Hotelling trace*, *Pillai's trace* y *Roy's largest root* aplicados al conjunto de medidas temporales. Comparación entre entrevistador automático y humano y las diferentes poblaciones.

	MANOVA							
	Var. Agrup. = Enfermedad (HC-AD)		**Var. Agrup. = Grado (HC-AD1)**		**Var. Agrup. = Grado (HC-AD2)**		**Var. Agrup. = Grado (AD1-AD2)**	
	p-value		***p-value***		***p-value***		***p-value***	
Estadístico/ Entrevist.	**Humano**	**Automático**	**Humano**	**Automático**	**Humano**	**Automático**	**Humano**	**Automático**
W	0,00	0,00	0,00	0,00	0,00	0,00	0,50	0,69
P	0,00	0,00	0,00	0,00	0,00	0,00	0,50	0,69
L	0,00	0,00	0,00	0,00	0,00	0,00	0,50	0,69
R	0,00	0,00	0,00	0,00	0,00	0,00	0,50	0,69

HC: sujeto de control, AD1: Alzheimer grado leve, AD2: Alzheimer grado moderado, *W*: *Wilks' lambda*, *P*: *Pillai's trace*, *L*: *Lawley-Hotelling trace*, *R*: *Roy's largest root.*

7.2 Temperatura Emocional

Para saber exactamente qué variables de Temperatura Emocional (descritas en el Capítulo 6) son discriminantes AD y cuáles no para cada entrevistador, se ha realizado un análisis estadístico descriptivo, un estudio paramétrico para ver si las muestras reflejan una distribución normal [243] y un estudio no paramétrico basado en tres tipos de pruebas: Test de Wilcoxon, Test de Kruskal-Wallis y Test de la Mediana. Por último se ha realizado un análisis multivariante MANOVA para estudiar el comportamiento de las cinco variables como conjunto.

7.2.1 Análisis de estadística descriptiva

Como se ha indicado previamente, en un primer paso se ha realizado un análisis de estadística descriptiva basado específicamente en los valores medio y de desviación estándar sobre cada una de las variables del conjunto de muestras.

En este apartado se hace una clasificación de las muestras atendiendo a cada una de las tres poblaciones bajo estudio (HC, AD1, AD2). En el caso de la población denominada AD, se incluyen las muestras relativas a los grados leves y moderados (grupos AD1 y AD2). A fin de facilitar la comparación entre los diferentes entrevistadores, en la Tabla 7-5 se muestran los valores de las variables obtenidos de cada uno de ellos. A simple vista, para algunas de las variables analizadas (como puede ser la variable te_d) puede observarse similitud entre los valores obtenidos del entrevistador humano y automático.

7.2.2 Análisis paramétrico

Los resultados obtenidos no siguen una distribución normal por lo que, en este caso, no estaría indicado realizar un análisis paramétrico. A partir de la regresión lineal de todas las variables temporales se observa que la recta de la regresión no se ajusta correctamente a todos los datos. Los resultados del test de normalidad de residuos, en concreto el test de *skewness* y curtosis, nos indica a partir de los valores de *Chi*

cuadrado que la regresión no es correcta y que los residuos no seguirían dicha distribución normal.

Tabla 7-5 Valores de estadística descriptiva de las medidas de Temperatura Emocional para cada población y para cada entrevistador: valor medio (μ) y desviación estándar (σ).

	Poblaciones							
	HC		**AD1**		**AD2**		**AD (AD1 + AD2)**	
Variable/ Entrevist.	**Humano** μ (σ)	**Automático** μ (σ)	**Humano** μ (σ)	**Automático** μ (σ)	**Humano** μ (σ)	**Automático** μ (σ)	**Humano** μ (σ)	**Automático** μ (σ)
te_d	52,92 (13,79)	52,68 (14,05)	57,49 (9,01)	57,37 (11,94)	48,91 (8,82)	57,13 (14,48)	56,12 (9,37)	57,29 (12,75)
$\overline{te_c}$	29,53 (29,4)	28,42 (27,76)	23,36 (29,84)	33,32 (29,84)	38,16 (26,41)	34,40 (29,89)	25,73 (28,47)	33,67 (29,72)
$\sigma^2_{te_c}$	451,50 (445,2)	435,72 (431,75)	390,53 (471,28)	514,74 (476,92)	767,26 (514,69)	583,80 (504,09)	487,93 (487,93)	537,14 (484,69)
$\tilde{\mu}_{te_{c_3}}$	-0,12 (0,51)	-0,12 (0,5)	-0,19 (0,38)	-0,29 (0,50)	0,01 (0,28)	-0,19 (0,53)	-0,16 (0,36)	-0,26 (0,50)
$Kurt_{te_c}$	2,42 (0,6)	2,37 (0,73)	2,23 (0,46)	2,36 (0,64)	1,93 (0,19)	2,24 (0,76)	2,18 (0,44)	2,32 (0,68)

HC: sujeto de control, AD1: Alzheimer grado leve, AD2: Alzheimer grado moderado.

7.2.3 Análisis no paramétrico

En la Tabla 7-6 se muestran los resultados obtenidos de las tres pruebas no paramétricas llevadas a cabo para comparar el entrevistador humano y el automático.

Los valores en los que la *Prob|z|* es mayor de 0,05 (*Prob|z| > 0,05*), casos en los que no hay diferencia al comparar las medidas de las poblaciones, se han representado en color gris. Estos valores avalarían la hipótesis nula y, por tanto, indicarían que una variable no es discriminante AD.

Puede observarse que para este tipo de variables no resulta tan evidente la discriminación entre poblaciones. Encontramos, de manera residual, algunos valores que nos permitirían rechazar la hipótesis nula (te_d y $\tilde{\mu}_{te_{c_3}}$, para el caso del entrevistador automático, y la variable $Kurt_{te_c}$ para el entrevistador humano).

7.2.4 Análisis multivariante MANOVA

Antes de realizar el análisis multivariante MANOVA, es necesario comprobar que las muestras siguen una distribución gaussiana. A tal efecto, hemos realizado una prueba de normalidad basada en asimetría y curtosis.

Una vez verificado que los resultados de la prueba efectivamente apoyan la aplicación de un análisis tipo MANOVA, los valores *p* obtenidos del análisis multivariado se muestran en la Tabla 7-7, donde *W*, *P*, *L* y *R* hacen referencia a los estadísticos *Wilks' lambda*, *Pillai's trace*, *Lawley-Hotelling trace* y *Roy's largest root*, respectivamente.

Al igual que en el análisis anterior, en gris, se muestran aquellos estadísticos que indicarían que el conjunto de variables no discrimina AD. Como se puede observar, los cuatro estadísticos indican que las variables dependientes bajo estudio resultan discriminantes únicamente bajo el escenario que enfrenta a las poblaciones HC y AD.

En base a estos resultados, aquellos casos en los que la variable de agrupamiento se basa en el grado de la enfermedad este conjunto de variables dejaría de ser discriminante AD.

Tabla 7-6 Resultados análisis univariante no paramétrico: qué variables de Temperatura Emocional son o no discriminantes comparando las diferentes poblaciones y entrevistadores.

	Test Wilcoxon		**Test Kruskal-Wallis**		**Test Mediana**	
	Prob|z|		**χ^2**		**Pearson χ^2**	
Variable/ Entrevist.	**Humano**	**Automático**	**Humano**	**Automático**	**Humano**	**Automático**
HC vs AD						
te_d	0,34	0,05	0.342	0,05	0,56	0,12
$\overline{te_c}$	0,61	0,19	0.634	0,21	0,68	0,18
$\sigma^2_{te_c}$	0,81	0,08	0.824	0,09	0,93	0,28
$\tilde{\mu}_{te_{c_3}}$	0,79	0,06	0.791	0,06	0,93	0,05
$Kurt_{te_c}$	0,13	0,55	0.129	0,55	0,37	0,96
HC vs AD1						
te_d	0,17	0,05	0,17	0,05	0,26	0,13
$\overline{te_c}$	0,47	0,28	0,50	0,30	0,66	0,23
$\sigma^2_{te_c}$	0,74	0,25	0,76	0,28	0,66	0,36
$\tilde{\mu}_{te_{c_3}}$	0,60	0,02	0,60	0,02	0,93	0,03
$Kurt_{te_c}$	0,34	0,93	0,34	0,93	0,66	0,36
HC vs AD2						
te_d	0,41	0,27	0,41	0,27	0,60	0,46
$\overline{te_c}$	0,69	0,30	0,71	0,32	0,60	0,57
$\sigma^2_{te_c}$	0,12	0,06	0,14	0,07	0,60	0,35
$\tilde{\mu}_{te_{c_3}}$	0,54	0,85	0,54	0.,8483	0,60	0,85
$Kurt_{te_c}$	0,05	0,16	0,05	0,16	0,12	0,35
AD1 vs AD2						
te_d	0,08	0,75	0,08	0,75	0,12	0,59
$\overline{te_c}$	0,52	0,87	0,55	0,87	0,53	0,89
$\sigma^2_{te_c}$	0,11	0,32	0,14	0,34	0,53	0,79
$\tilde{\mu}_{te_{c_3}}$	0,18	0,20	0,18	0,20	0,65	0,28
$Kurt_{te_c}$	0,24	0,14	0,24	0,14	0,12	0,08

HC: sujeto de control, AD1: Alzheimer grado leve, AD2: Alzheimer grado moderado

Tabla 7-7 Resultados del análisis multivariante MANOVA para los estadísticos *Wilks' lambda*, *Lawley-Hotelling trace*, *Pillai's trace* y *Roy's largest root* aplicados al conjunto de medidas de Temperatura Emocional. Comparación entre entrevistador automático y humano y las diferentes poblaciones.

	MANOVA							
	Var. Agrup. = Enfermedad (HC-AD)		**Var. Agrup. = Grado (HC-AD1)**		**Var. Agrup. = Grado (HC-AD2)**		**Var. Agrup. = Grado (AD1-AD2)**	
	p-value		***p-value***		***p-value***		***p-value***	
Estadístico/ Entrevist.	**Humano**	**Automático**	**Humano**	**Automático**	**Humano**	**Automático**	**Humano**	**Automático**
W	0,14	0,01	0,12	0,05	0,40	0,06	0,35	0,74
P	0,14	0,01	0,12	0,05	0,40	0,06	0,35	0,74
L	0,14	0,01	0,12	0,05	0,40	0,06	0,35	0,74
R	0,14	0,01	0,12	0,05	0,40	0,06	0,35	0,74

HC: sujeto de control, AD1: Alzheimer grado leve, AD2: Alzheimer grado moderado, *W*: *Wilks' lambda*, *P*: *Pillai's trace*, *L*: *Lawley-Hotelling trace*, *R*: *Roy's largest root.*

7.3 Análisis multivariante MANOVA: medidas temporales y Temperatura Emocional

Como se ha hecho en los casos anteriores, antes de realizar el análisis multivariante MANOVA, es necesario comprobar que las muestras siguen una distribución gaussiana. A tal efecto, hemos realizado una prueba de normalidad basada en asimetría y curtosis.

Una vez verificado que los resultados de la prueba efectivamente apoyan la aplicación de un análisis tipo MANOVA, los valores *p* obtenidos del análisis multivariado se muestran en la Tabla 7-8, donde *W*, *P*, *L* y *R* hacen referencia a los estadísticos *Wilks' lambda*, *Pillai's trace*, *Lawley-Hotelling trace* y *Roy's largest root*, respectivamente.

En este caso, los resultados del análisis MANOVA, cuando se utilizan las diez características temporales y de Temperatura Emocional en conjunto, indican que dichas variables son discriminantes excepto para el caso en el que se enfrentan las poblaciones AD1 y AD2.

Tabla 7-8 Resultados del análisis multivariante MANOVA para los estadísticos *Wilks' lambda*, *Lawley-Hotelling trace*, *Pillai's trace* y *Roy's largest root* aplicados al conjunto completo de medidas temporales y de Temperatura Emocional. Comparación entre entrevistador automático y humano y los diferentes clasificadores.

	MANOVA							
	Var. Agrup. = Enfermedad (HC-AD)		**Var. Agrup. = Grado (HC-AD1)**		**Var. Agrup. = Grado (HC-AD2)**		**Var. Agrup. = Grado (AD1-AD2)**	
	p-value		***p-value***		***p-value***		***p-value***	
Estadístico/ Entrevist.	**Humano**	**Automático**	**Humano**	**Automático**	**Humano**	**Automático**	**Humano**	**Automático**
W	0,00	0,00	0,00	0,00	0,00	0,00	0,77	0,84
P	0,00	0,00	0,00	0,00	0,00	0,00	0,77	0,84
L	0,00	0,00	0,00	0,00	0,00	0,00	0,77	0,84
R	0,00	0,00	0,00	0,00	0,00	0,00	0,77	0,84

HC: sujeto de control, AD1: Alzheimer grado leve, AD2: Alzheimer grado moderado, *W*: *Wilks' lambda*, *P*: *Pillai's trace*, *L*: *Lawley-Hotelling trace*, *R*: *Roy's largest root.*

Capítulo 8 Clasificadores

8.1 Análisis discriminante mediante el *software* de análisis estadístico Stata

Con el fin de comparar el rendimiento del entrevistador humano y automático, en este apartado se estudian los clasificadores implementados en el *software* de análisis estadístico Stata. Se realiza el mismo proceso tanto para el set de medidas temporales y de carga emocional, como para la combinación de ambos tipos de características.

8.1.1 Medidas temporales

En este apartado se van a tener en cuenta las cinco medidas temporales MediaHabla ($\overline{t_s}$), VarHabla ($\sigma_{t_s}^2$), SKWHabla ($\tilde{\mu}_{t_{s_3}}$), KRTHabla ($Kurt_{t_s}$) e INDHabla (Ind_{t_s}) definidas. En base a estas cinco variables y a la aplicación de los clasificadores LDA, logístico y kNN (para *n = 1*, *n = 3*, *n = 5*) con configuración *leave-one-out* sobre las muestras de la base de datos, realizamos, a su vez, dos tipos de clasificaciones. La primera basada en la presencia o ausencia de enfermedad (HC: sujeto de control y AD: población con Alzheimer) y la segunda basada en tres poblaciones diferentes según el grado de enfermedad (AD1: grado leve y AD2: grado moderado).

Clasificación multivariante basada en presencia o ausencia de enfermedad

La Tabla 8-1 muestra los resultados obtenidos representados en una matriz de confusión para cada entrevistador y clasificador. En ella se muestra información sobre el número de muestras clasificadas y su porcentaje del total.

De la misma forma, los valores de tasa de éxito, sensibilidad y especificidad de cada entrevistador se muestran a continuación en la Tabla 8-2, donde podemos ver que los mejores resultados se obtienen para los clasificadores LDA y logístico, para el entrevistador humano y automático, respectivamente.

Tabla 8-1 Análisis discriminante: matriz de confusión para la clasificación multivariante LDA, clasificador logístico y kNN en base a presencia (1) o ausencia de enfermedad (0). Medidas temporales.

		Entrevistador automático			Entrevistador humano		
Clasificador	**True Enfermedad**	**0**	**1**	**Total**	**0**	**1**	**Total**
LDA	**0**	121 (87,68%)	17 (12,32%)	138 (100%)	40 (86,96)	6 (13,04)	46 (100%)
	1	31 (27,68%)	81 (72,32%)	112 (100%)	4 (15,38)	22 (84,62%)	26 (100%)
	Total	152 (60,80%)	98 (39,20%)	250 (100%)	44 (61,11)	28 (38,89)	72 (100%)
Logístico	**0**	117 (84,78%)	21 (15,22%)	138 (100%)	43 (93,48%)	3 (6,52%)	46 (100%)
	1	29 (25,89%)	83 (74,11%)	112 (100%)	2 (7,69%)	24 (92,31%)	26 (100%)
	Total	146 (58,40%)	104 (41,60%)	250 (100%)	45 (62,50%)	27 (37,50%)	72 (100%)
kNN (n=1)	**0**	84 (60,87%)	54 (39,13%)	138 (100%)	32 (69,57%)	14 (30,43%)	46 (100%)
	1	41 (36,61%)	71 (63,39%)	112 (100%)	15 (57,69%)	11 (42,31%)	26 (100%)
	Total	125 (50%)	125 (50%)	250 (100%)	47 (65,28%)	25 (34,72%)	72 (100%)
kNN (n=3)	**0**	97 (70,29%)	41 (29,71%)	138 (100%)	30 (65,22%)	16 (34,78%)	46 (100%)
	1	50 (44,64%)	62 (55,36%)	112 (100%)	14 (53,85%)	12 (46,15%)	26 (100%)
	Total	147 (58,80%)	103 (41,20%)	250 (100%)	44 (51,11%)	28 (38,89%)	72 (100%)
kNN (n=5)	**0**	98 (71,01%)	40 (28,99%)	138 (100%)	21 (45,65%)	25 (54,35%)	46 (100%)
	1	51 (45,54%)	61 (54,46%)	112 (100%)	6 (23,08%)	20 (76,92%)	26 (100%)
	Total	149 (59,60%)	101 (40,40%)	250 (100%)	27 (37,50%)	45 (62,50%)	72 (100%)

Tabla 8-2 Análisis discriminante: valores de tasa de éxito, sensibilidad y especificidad para entrevistador automático y humano en base a clasificación multivariante LDA, clasificador logístico y kNN por ausencia o presencia de enfermedad. Medidas temporales.

	Clasificador	**Tasa de éxito [%]**	**Sensibilidad [%]**	**Especificidad [%]**
Entrevistador automático	**LDA**	80,00%	88,00%	59,00%
	Logístico	79,40%	85,00%	75,00%
	kNN (n=1)	62,13%	61,00%	64,00%
	kNN (n=3)	62,82%	70,00%	55,00%
	kNN (n=5)	62,74%	72,00%	55,00%
Entrevistador humano	**LDA**	85,80%	87,00%	84,00%
	Logístico	92,90%	94,00%	93,00%
	kNN (n=1)	63,60%	70,00%	42,00%
	kNN (n=3)	55,70%	66,00%	46,00%
	kNN (n=5)	61,29%	45,00%	76,00%

Clasificación multivariante basada en diferentes grados de enfermedad

Los resultados obtenidos al clasificar las muestras en diferentes grados de enfermedad: ausencia de enfermedad (0), leve (1) y moderada (2), divididos por tipo de entrevistador y clasificador utilizado, se muestran en la Tabla 8-3.

De igual forma, la Tabla 8-4 muestra los valores de tasa de éxito, sensibilidad y especificidad obtenidos a partir de la Tabla 8-3 para cada uno de los entrevistadores y clasificadores. Podemos ver que los mejores resultados de clasificación se logran, para ambos entrevistadores, con los clasificadores logísticos.

Tabla 8-3 Análisis discriminante: matriz de confusión para la clasificación multivariante LDA, clasificador logístico y kNN en función de los diferentes grados de la enfermedad: ausencia de enfermedad (0), leve (1) y moderada (2). Medidas temporales.

Clasif.	True Grado	Entrevistador automático				Entrevistador humano			
		0	**1**	**2**	**Total**	**0**	**1**	**2**	**Total**
LDA	**0**	116 (84,06%)	14 (10,14%)	8 (5,80%)	138 (100%)	40 (86,96%)	6 (13,04%)	0 (0%)	46 (100%)
	1	20 (26,67%)	27 (36,00%)	28 (37,33%)	75 (100%)	3 (14,29%)	15 (71,43%)	3 (14,29%)	21 (100%)
	2	10 (27,03%)	8 (21,62%)	19 (51,35)	37 (100%)	1 (20%)	2 (40%)	2 (40%)	5 (100%)
	Total	146 (58,40%)	49 (19,60%)	55 (22%)	250 (100%)	44 (61,11%)	23 (31,94%)	5 (6,94%)	72 (100%)
Logístico	**0**	116 (84,06%)	14 (10,14%)	8 (5,80%)	138 (100%)	42 (91,3%)	4 (8,70%)	0 (0%)	46 (100%)
	1	17 (22,67%)	32 (42,67%)	26 (34,67%)	75 (100%)	1 (4,76%)	17 (80,95%)	3 (14,29%)	21 (100%)
	2	9 (24,32%)	5 (13,51%)	23 (62,16%)	37 (100%)	1 (20%)	0 (0%)	4 (80%)	5 (100%)
	Total	142 (56,8%)	51 (20,40%)	57 (22,80%)	250 (100%)	44 (61,11%)	21 (29,17%)	4 (80%)	72 (100%)
kNN (n=1)	**0**	84 (60,87%)	38 (27,54%)	16 (11,59%)	138 (100%)	32 (69,57%)	12 (26,09%)	2 (4,35%)	46 (100%)
	1	28 (37,33%)	33 (44%)	14 (18,67%)	75 (100%)	12 (57,14%)	7 (33,33%)	2 (9,52%)	21 (100%)
	2	13 (35,14%)	14 (37,84%)	10 (27,03%)	37 (100%)	3 (60%)	2 (40%)	0 (0%)	5 (100%)
	Total	125 (50%)	85 (34%)	40 (16 %)	250 (100%)	47 (65,28%)	21 (29,17%)	4 (5,56%)	72 (100%)
kNN (n=3)	**0**	71 (51,45%)	20 (14,49%)	47 (34,06%)	138 (100%)	17 (36,96%)	21 (45,65%)	8 (17,39%)	46 (100%)
	1	31 (41,33%)	14 (18,67%)	21 (28%)	75 (100%)	4 (19,05%)	9 (42,86%)	8 (38,1%)	21 (100%)
	2	10 (27,03%)	12 (32,43%)	15 (40,54%)	37 (100%)	2 (40%)	3 (60%)	0 (0%)	5 (100%)
	Total	112 (44,80%)	46 (18,40%)	83 (33,2%)	250 (100%)	23 (31,94%)	33 (45,83%)	16 (22,22%)	72 (100%)
kNN (n=5)	**0**	59 (42,75%)	36 (26,09)	43 (31,16%)	138 (100%)	18 (39,13%)	15 (32,61%)	13 (28,26%)	46 (100%)
	1	13 (17,33%)	27 (36%)	28 (37,33%)	75 (100%)	3 (14,29%)	7 (33,33%)	11 (52,38%)	21 (100%)
	2	7 (18,92%)	15 (40,54%)	15 (40,54%)	37 (100%)	2 (40%)	3 (60%)	0 (0%)	5 (100%)
	Total	79 (31,6%)	78 (31,2%)	86 (34,4%)	250 (100%)	23 (31,94%)	25 (34,72%)	24 (33,33%)	72 (100%)

Tabla 8-4 Análisis discriminante: valores de tasa de éxito, sensibilidad y especificidad para entrevistador automático y humano basado en clasificación multivariante LDA, clasificador logístico y kNN por grados de la enfermedad. Medidas temporales.

	Clasificador	Tasa de éxito [%]	Sensibilidad [%]	Especificidad [%]
Entrevistador automático	**LDA**	78,63%	76,40%	83,00%
	Logístico	80,42%	78,50%	83,50%
	kNN (n=1)	62,13%	62,50%	62,00%
	kNN (n=3)	55,80%	56,00%	55,00%
	kNN (n=5)	61,85%	69,00%	58,00%
Entrevistador humano	**LDA**	85,79%	85,14%	86,60%
	Logístico	91,80%	92,80%	92,00%
	kNN (n=1)	56,00%	55,00%	58,00%
	kNN (n=3)	50,00%	50,00%	50,00%
	kNN (n=5)	62,95%	67,20%	59,00%

8.1.2 Temperatura Emocional

De la misma manera que en el apartado anterior, se van a tener en cuenta las medidas, en este caso, de Temperatura Emocional: TEd (te_d), MediaTEc ($\overline{te_c}$), VarTEc ($\sigma^2_{te_c}$), SKWTEc ($\tilde{\mu}_{te_{c_3}}$) y KRTTEc ($Kurt_{te_c}$).

En base a estas cinco variables y a la aplicación de clasificadores LDA, logístico y kNN (para *n = 1, n = 3, n = 5*) con configuración *leave-one-out* sobre las muestras del entrevistador automático y humano, realizamos, a su vez, dos tipos de clasificaciones. La primera basada en la presencia o ausencia de enfermedad y la segunda basada en tres poblaciones diferentes según el grado de enfermedad (AD leve o moderada).

Clasificación multivariante basada en presencia o ausencia de enfermedad

La Tabla 8-5 muestra los resultados obtenidos representados en una matriz de confusión para cada entrevistador y clasificador. En ella se muestra información sobre el número de muestras clasificadas y su porcentaje sobre el total.

Tabla 8-5 Análisis discriminante: matriz de confusión para la clasificación multivariante LDA, clasificador logístico y kNN en base a presencia (1) o ausencia de enfermedad (0). Temperatura Emocional.

		Entrevistador automático			Entrevistador humano		
Clasificador	**True Enfermedad**	**0**	**1**	**Total**	**0**	**1**	**Total**
LDA	**0**	75 (54,35%)	63 (45,65%)	138 (100%)	22 (47,83%)	24 (52,17%)	46 (100%)
	1	47 (42,34%)	64 (57,66%)	111 (100%)	13 (52,00%)	12 (48,00%)	25 (100%)
	Total	122 (49,00%)	127 (51%)	249 (100%)	35 (49,30%)	36 (50,70%)	71 (100%)
Logístico	**0**	81 (58,70%)	57 (41,30%)	138 (100%)	26 (56,52%)	20 (43,48%)	46 (100%)
	1	44 (39,64%)	67 (60,36%)	111 (100%)	10 (40,00%)	15 (60,00%)	25 (100%)
	Total	125 (50,20%)	124 (49,80%)	249 (100%)	36 (50,70%)	35 (49,30%)	71 (100%)
kNN (n=1)	**0**	78 (56,52%)	60 (43,48%)	138 (100%)	27 (58,70%)	19 (41,30%)	46 (100%)
	1	59 (53,15%)	52 (46,85%)	111 (100%)	18 (72,00%)	7 (28,00%)	25 (100%)
	Total	137 (55,02%)	112 (44,98%)	249 (100%)	45 (63,38%)	26 (36,62%)	71 (100%)
kNN (n=3)	**0**	76 (55,07%)	62 (44,93%)	138 (100%)	32 (69,57%)	14 (30,43%)	46 (100%)
	1	67 (60,36%)	44 (39,64%)	111 (100%)	19 (76,00%)	6 (24,00%)	25 (100%)
	Total	143 (57,43%)	106 (42,57%)	249 (100%)	51 (71,83%)	20 (28,17%)	71 (100%)
kNN (n=5)	**0**	86 (62,32%)	52 (37,68%)	138 (100%)	16 (34,78%)	30 (65,22%)	46 (100%)
	1	62 (55,86%)	49 (44,14%)	111 (100%)	10 (40,00%)	15 (60,00%)	25 (100%)
	Total	148 (59,44%)	101 (40,56%)	249 (100%)	26 (36,62%)	45 (63,38)	71 (100%)

A partir de estos resultados, los valores de tasa de éxito, sensibilidad y especificidad de cada entrevistador se muestran a continuación en la Tabla 8-6. En ella podemos ver que, si bien el rendimiento es inferior al obtenido por las variables temporales, los mejores resultados se obtienen para los clasificadores logísticos, tanto para el entrevistador humano como para el automático.

Clasificación multivariante basada en diferentes grados de enfermedad

Los resultados obtenidos al clasificar las muestras en diferentes grados de enfermedad: ausencia de enfermedad (0), leve (1) y moderada (2), divididos por tipo de entrevistador y clasificador utilizado, se muestran en la Tabla 8-7.

Tabla 8-6 Análisis discriminante: valores de tasa de éxito, sensibilidad y especificidad para entrevistador automático y humano en base a clasificación multivariante LDA, clasificador logístico y kNN por ausencia o presencia de enfermedad. Temperatura Emocional.

	Clasificador	Tasa de éxito [%]	Sensibilidad [%]	Especificidad [%]
Entrevistador automático	**LDA**	55,82%	57,66%	54,35%
	Logístico	59,44%	60,36%	58,70%
	kNN (n=1)	52,21%	46,85%	56,52%
	kNN (n=3)	48,19%	39,64%	55,07%
	kNN (n=5)	54,22%	44,14%	62,32%
Entrevistador humano	**LDA**	47,89%	48,00%	47,83%
	Logístico	57,75%	60,00%	56,52%
	kNN (n=1)	47,89%	28,00%	58,70%
	kNN (n=3)	53,52%	24,00%	69,57%
	kNN (n=5)	43,66%	60,00%	34,78%

Tabla 8-7 Análisis discriminante: matriz de confusión para la clasificación multivariante LDA, clasificador logístico y kNN en función de los diferentes grados de la enfermedad: ausencia de enfermedad (0), leve (1) y moderada (2). Temperatura Emocional.

Clasif.	True Grado	Entrevistador automático				Entrevistador humano			
		0	**1**	**2**	**Total**	**0**	**1**	**2**	**Total**
LDA	**0**	64 (46.38%)	34 (24.64%)	40 (28.99%)	138 (100%)	18 (39.13%)	14 (30.43%)	14 (30.43%)	46 (100%)
	1	28 (37.33%)	18 (24.00%)	29 (38.67%)	75 (100%)	10 (47.62%)	8 (38.10%)	3 (14.29%)	21 (100%)
	2	12 (33,33%)	10 (27.78%)	14 (38.89%)	36 (100%)	1 (25.00%)	0 (0%)	3 (75,00%)	4 (100%)
	Total	104 (41.77%)	62 (24.90%)	83 (33.33%)	249 (100%)	29 (40.85%)	22 (30.99%)	20 (28.17%)	71 (100%)
Logístico	**0**	70 (50.72%)	30 (21.74%)	38 (27.54%)	138 (100%)	25 (54.35%)	15 (32.61%)	6 (13.04%)	46 (100%)
	1	22 (29.33%)	31 (41.33%)	22 (29.33%)	75 (100%)	7 (33.33%)	11 (52.38%)	3 (14.29%)	21 (100%)
	2	12 (33.33%)	9 (25.00%)	15 (41.67%)	36 (100%)	0 (0%)	0 (0%)	4 (100%)	4 (100%)
	Total	104 (41.77%)	70 (28.11%)	75 (30.12%)	249 (100%)	32 (45.07%)	26 (36.62%)	13 (18.31%)	71 (100%)
kNN (n=1)	**0**	78 (56.52%)	38 (27.54%)	22 (15.94%)	138 (100%)	27 (58.70%)	13 (39.13%)	1 (2.17%)	46 (100%)
	1	42 (56.00%)	21 (28.00%)	12 (16.00%)	75 (100%)	15 (71.43%)	5 (23.81%)	1 (4.76%)	21 (100%)
	2	17 (47.22%)	9 (25,00)	10 (27.78%)	36 (100%)	3 (75.00%)	1 (25.00%)	0 (0%)	4 (100%)
	Total	137 (55.02%)	68 (27.31%)	44 (17.67%)	249 (100%)	45 (63.38%)	24 (33.80%)	2 (2.82%)	71 (100%)
kNN (n=3)	**0**	59 (42.75%)	28 (20.29%)	51 (36.96%)	138 (100%)	9 (19.57%)	27 (58.70%)	10 (21.74%)	46 (100%)
	1	34 (45.33%)	17 (22.67%)	24 (32.00%)	75 (100%)	7 (33.33%)	9 (42.86%)	5 (23.81%)	21 (100%)
	2	16 (44.44%)	5 (13.89%)	15 (41.67%)	36 (100%)	0 (0%)	4 (100%)	0 (0%)	4 (100%)
	Total	109 (43.78%)	50 (20.08%)	90 (36.14%)	249 (100%)	16 (22.54%)	40 (56.34%)	15 (21.13%)	71 (100%)
kNN (n=5)	**0**	36 (26.09%)	54 (39.13%)	48 (34.78%)	138 (100%)	14 (30.43%)	17 (36.96%)	15 (32.61%)	46 (100%)
	1	21 (28.00%)	29 (38.67%)	25 (33.33%)	75 (100%)	7 (33.33%)	7 (33.33%)	7 (33.33%)	21 (100%)
	2	8 (22.22%)	6 (16.67%)	22 (61.11%)	36 (100%)	1 (25.00%)	2 (50.00%)	1 (25.00%)	4 (100%)
	Total	65 (26.10%)	89 (35.74%)	95 (38.15%)	249 (100%)	22 (30.99%)	26 (36.62%)	23 (32.39%)	71 (100%)

La Tabla 8-8 muestra los valores de tasa de éxito, sensibilidad y especificidad obtenidos para cada uno de los entrevistadores y clasificadores para la clasificación multivariante basada en diferentes grados de enfermedad. De nuevo, podemos ver que los mejores resultados se logran para los clasificadores logísticos sea cual sea el entrevistador utilizado.

Tabla 8-8 Análisis discriminante: valores de tasa de éxito, sensibilidad y especificidad para entrevistador automático y humano basado en clasificación multivariante LDA, clasificador logístico y kNN por grados de la enfermedad. Temperatura Emocional.

	Clasificador	Tasa de éxito [%]	Sensibilidad [%]	Especificidad [%]
Entrev. autom.	**LDA**	38,55%	63,96%	46,38%
	Logístico	46,59%	69,37%	50,72%
	kNN (n=1)	43,78%	46,85%	56,52%
	kNN (n=3)	36,55%	54,95%	42,75%
	kNN (n=5)	34,94%	73,87%	26,09%
Entrev. humano	**LDA**	40,85%	56,00%	39,13%
	Logístico	56,34%	72,00%	54,35%
	kNN (n=1)	48,48%	28,00%	65,85%
	kNN (n=3)	25,35%	72,00%	19,57%
	kNN (n=5)	30,99%	68,00%	30,43%

8.1.3 Clasificación de la combinación de medidas temporales y de Temperatura Emocional

En este apartado se van a tener en cuenta las medidas temporales y de Temperatura Emocional: MediaHabla ($\overline{t_S}$), VarHabla ($\sigma^2_{t_S}$), SKWHabla ($\tilde{\mu}_{t_{S_3}}$), KRTHabla ($Kurt_{t_S}$), INDHabla (Ind_{t_S}), TEd (te_d), MediaTEc ($\overline{te_c}$), VarTEc ($\sigma^2_{te_c}$), SKWTEc ($\tilde{\mu}_{te_{c_3}}$) y KRTTEc ($Kurt_{te_c}$). En base a estas diez variables y a la aplicación de los clasificadores LDA, logístico y kNN (para *n = 1*, *n = 3*, *n = 5*) con configuración *leave-one-out* sobre las muestras del entrevistador automático y humano, realizamos, a su vez, dos tipos de clasificaciones. La primera basada en la presencia o ausencia de enfermedad y la segunda basada en tres poblaciones diferentes según el grado de enfermedad (AD leve y moderada).

Clasificación multivariante basada en presencia o ausencia de enfermedad

La Tabla 8-9 muestra los resultados obtenidos representados en una matriz de confusión para cada entrevistador y clasificador. En ella se muestra información sobre el número de muestras clasificadas y su porcentaje sobre el total.

Los valores de tasa de éxito, sensibilidad y especificidad para cada entrevistador se muestran a continuación en la Tabla 8-10. A partir de ella podemos ver que los mejores resultados se obtienen con los clasificadores LDA y logístico, para el entrevistador automático y humano, respectivamente.

Tabla 8-9 Análisis discriminante: matriz de confusión para la clasificación multivariante LDA, clasificador logístico y kNN en base a presencia (1) o ausencia de enfermedad (0). Medidas temporales y Temperatura Emocional.

Clasificador	True Enferm.	Entrevistador automático			Entrevistador humano		
		0	1	Total	0	1	Total
LDA	0	120 (86.96%)	18 (13.04%)	138 (100%)	42 (91.30%)	4 (8.70%)	46 (100%)
	1	33 (29.73%)	78 (70.27%)	111 (100%)	4 (16.00%)	21 (84.00%)	25 (100%)
	Total	153 (61.45%)	96 (38.55%)	249 (100%)	46 (64.79%)	25 (35.21%)	71 (100%)
Logístico	0	119 (86.23%)	19 (13.77%)	138 (100%)	43 (93.48%)	3 (6.52%)	46 (100%)
	1	29 (26.13%)	82 (73.87%)	111 (100%)	2 (8.00%)	23 (92.00%)	25 (100%)
	Total	148 (59.44%)	101 (40.56%)	249 (100%)	45 (63.38%)	26 (36.62%)	71 (100%)
kNN (n=1)	0	85 (61.59%)	53 (38.41%)	138 (100%)	26 (56.52%)	20 (43.48%)	46 (100%)
	1	58 (52.25%)	53 (47.75%)	111 (100%)	15 (60.00%)	10 (40.00%)	25 (100%)
	Total	143 (57.43%)	106 (42.57%)	249 (100%)	41 (57.75%)	30 (42.25%)	71 (100%)
kNN (n=3)	0	78 (56.52%)	60 (43.48%)	138 (100%)	30 (65.22%)	16 (34.78%)	46 (100%)
	1	67 (60.36%)	44 (39.64%)	111 (100%)	19 (76.00%)	6 (24.00%)	25 (100%)
	Total	145 (58.23%)	104 (41.77%)	249 (100%)	49 (69.01%)	22 (30.99%)	71 (100%)
kNN (n=5)	0	84 (60.87%)	54 (39.13%)	138 (100%)	11 (23.91%)	35 (76.09%)	46 (100%)
	1	61 (54.95%)	50 (45.05%)	111 (100%)	7 (28.00%)	18 (72.00%)	25 (100%)
	Total	145 (58.23%)	104 (41.77%)	249 (100%)	18 (25.35%)	53 (74.65%)	71 (100%)

Tabla 8-10 Análisis discriminante: valores de tasa de éxito, sensibilidad y especificidad para entrevistador automático y humano en base a clasificación multivariante LDA, clasificador logístico y kNN por ausencia o presencia de enfermedad. Medidas temporales y Temperatura Emocional.

	Clasificador	Tasa de éxito [%]	Sensibilidad [%]	Especificidad [%]
Entrevistador automático	LDA	79,52%	70,27%	86,96%
	Logístico	80,72%	73,87%	86,23%
	kNN (n=1)	55,42%	47,75%	61,59%
	kNN (n=3)	49,00%	39,64%	56,52%
	kNN (n=5)	53,82%	45,05%	60,87%
Entrevistador humano	LDA	88,73%	84,00%	91,30%
	Logístico	92,96%	92,00%	93,48%
	kNN (n=1)	50,70%	40,00%	56,52%
	kNN (n=3)	50,70%	24,00%	65,22%
	kNN (n=5)	40,85%	72,00%	23,91%

Clasificación multivariante basada en diferentes grados de enfermedad

Los resultados obtenidos al clasificar las muestras en diferentes grados de la enfermedad: ausencia de enfermedad (0), leve (1) y moderada (2), divididos por tipo de entrevistador y clasificador utilizado, se muestran en la Tabla 8-11.

La Tabla 8-12 presenta los valores de tasa de éxito, sensibilidad y especificidad obtenidos para cada uno de los entrevistadores y clasificadores. De nuevo, podemos ver que los mejores resultados se logran para los clasificadores logísticos para ambos entrevistadores.

Tabla 8-11 Análisis discriminante: matriz de confusión para la clasificación multivariante LDA, clasificador logístico y kNN en función de los diferentes grados de la enfermedad: ausencia de enfermedad (0), leve (1) y moderada (2). Medidas temporales y Temperatura Emocional.

Clasif.	True Grado	Entrevistador automático				Entrevistador humano			
		0	**1**	**2**	**Total**	**0**	**1**	**2**	**Total**
LDA	**0**	118 (85.51%)	14 (10.14%)	6 (4.35%)	138 (100%)	49 (86.96%)	4 (8.70%)	2 (4.35%)	46 (100%)
	1	22 (29.33%)	23 (30.67%)	30 (40.00%)	75 (100%)	3 (14.29%)	11 (52.38%)	7 (33.33%)	21 (100%)
	2	10 (27.78%)	7 (19.44%)	19 (52.78%)	36 (100%)	25 (25.00%)	2 (50.00%)	1 (25.00%)	4 (100%)
	Total	150 (60.24%)	44 (17.67%)	55 (22.09%)	249 (100%)	44 (61.97%)	17 (23.94%)	10 (14.08%)	71 (100%)
Logístico	**0**	115 (83.33%)	13 (9.42%)	10 (7.25%)	138 (100%)	45 (97.83%)	1 (2.17%)	0 (0%)	46 (100%)
	1	16 (21.33%)	35 (46.67%)	24 (32.00%)	75 (100%)	1 (4.76%)	20 (95.24%)	0 (0%)	21 (100%)
	2	10 (27.78%)	5 (13.89%)	21 (58.33%)	36 (100%)	0 (0%)	0 (0%)	4 (100%)	4 (100%)
	Total	141 (56.63%)	53 (21.29%)	55 (22.09%)	249 (100%)	46 (64.79%)	21 (29.58%)	4 (5.63%)	71 (100%)
kNN (n=1)	**0**	85 (61.59%)	35 (25.36%)	18 (13.04%)	138 (100%)	26 (56.52%)	17 (36.96%)	3 (6.52%)	46 (100%)
	1	40 (53.33%)	26 (34.67%)	9 (12.00%)	75 (100%)	13 (61.90%)	6 (28.57%)	2 (9.52%)	21 (100%)
	2	18 (50.00%)	9 (25.00%)	9 (25.00%)	36 (100%)	2 (50.00%)	2 (50.00%)	0 (0%)	4 (100%)
	Total	143 (57.43%)	70 (28.11%)	36 (14.46%)	249 (100%)	41 (57.75%)	25 (35.21%)	5 (7.04%)	71 (100%)
kNN (n=3)	**0**	65 (47.10%)	30 (21.74%)	43 (31.16%)	138 (100%)	10 (21.74%)	26 (56.52%)	10 (21.74%)	46 (100%)
	1	35 (46.67%)	20 (26.67%)	20 (26.67%)	75 (100%)	6 (28.57%)	10 (47.62%)	5 (23.81%)	21 (100%)
	2	18 (50.00%)	5 (13.89%)	13 (36.11%)	36 (100%)	0 (0%)	4 (100.00%)	0 (0%)	4 (100%)
	Total	118 (47.39%)	55 (22.09%)	76 (30.52%)	249 (100%)	16 (22.54%)	40 (56.34%)	15 (21.13%)	71 (100%)
kNN (n=5)	**0**	41 (29.71%)	52 (37.68%)	45 (32.61%)	138 (100%)	9 (19.57%)	22 (47.83%)	15 (32.61%)	46 (100%)
	1	19 (25.33%)	31 (41.33%)	25 (33.33%)	75 (100%)	5 (23.81%)	9 (42.86%)	7 (33.33%)	21 (100%)
	2	6 (16.67%)	8 (22.22%)	22 (61.11%)	36 (100%)	1 (25.00%)	2 (50.00%)	1 (25.00%)	4 (100%)
	Total	66 (26.51%)	91 (36.55%)	92 (36.95%)	249 (100%)	15 (21.13%)	33 (46.48%)	23 (32.39%)	71 (100%)

Tabla 8-12 Análisis discriminante: valores de tasa de éxito, sensibilidad y especificidad para entrevistador automático y humano basado en clasificación multivariante LDA, clasificador logístico y kNN por grados de la enfermedad. Medidas temporales y Temperatura Emocional.

	Clasificador	Tasa de éxito [%]	Sensibilidad [%]	Especificidad [%]
Entrevistador automático	**LDA**	64,26%	71,17%	85,51%
	Logístico	68,67%	76,58%	83,33%
	kNN(n=1)	48,19%	47,75%	61,59%
	kNN(n=3)	39,36%	52,25%	47,10%
	kNN(n=5)	37,75%	77,48%	29,71%
Entrevistador humano	**LDA**	58,65%	42,86%	89,09%
	Logístico	97,18%	96,00%	97,83%
	kNN(n=1)	45,07%	40,00%	56,52%
	kNN(n=3)	28,17%	76,00%	21,74%
	kNN(n=5)	26,76%	76,00%	19,57%

8.2 Modelos de clasificación mediante el sistema de cómputo numérico Matlab

8.2.1 Selección de características

A continuación se muestran los resultados obtenidos de aplicar la función *fscnca* desarrollada en Matlab (también conocida como selección de características mediante el análisis de componentes de vecindad para la clasificación) tanto a las características temporales, como a las de Temperatura Emocional. Esta función asigna un peso a cada característica dentro del conjunto mediante el uso de una adaptación diagonal del análisis de componentes de vecindad (NCA).

Para conocer la relevancia de cada una de las variables, se ha comprobado la afección que pueda tener el tipo de entrevistador empleado, así como la influencia del número de clases utilizadas: enfermedad o grados. Sin embargo, en ambos casos, los resultados de la selección de características se han mantenido relativamente constantes. En la Tabla 8-13 se muestra cómo han variado sus valores atendiendo a las condiciones descritas en el este párrafo.

Tabla 8-13 Selección de características: relevancia de las características temporales y de Temperatura Emocional a partir de la función *fscnca* de Matlab (A: entrevistador automático, H: entrevistador humano).

		$\bar{t_S}$	$\sigma^2_{t_S}$	$\tilde{\mu}_{t_{S_3}}$	$Kurt_{t_s}$	Ind_{t_s}	te_d	$\overline{te_c}$	$\sigma^2_{te_c}$	$\tilde{\mu}_{te_{c3}}$	$Kurt_{te_c}$
Enfermedad	**A**	0	0,9	0	1,6	4,1	0,9	2,1	0,15	0	0
	H	0	1,45	0	1,3	0	0,55	1	0,35	0	0
Grado	**A**	0	0	0,95	0	3,3	0,4	1,7	0,1	0	0
	H	0	1,3	0	1,25	0	0,45	0,95	0,3	0	0

A continuación, en la Tabla 8-14 se muestran, en concreto, los resultados del proceso de selección de características temporales tomando como referencia el conjunto de muestras del entrevistador automático y para el caso de que la clasificación sea biclase. Se acompaña, asimismo, de una representación gráfica de los mismos (Figura 8-1) en la que se puede apreciar el peso de cada una de ellas sobre el conjunto.

Tabla 8-14 Selección de características: resultados de la selección de medidas temporales.

Medida	Relevancia
$\bar{t_S}$	0
$\sigma^2_{t_S}$	0,9
$\tilde{\mu}_{t_{S_3}}$	0
$Kurt_{t_s}$	1,6
Ind_{t_s}	4,1

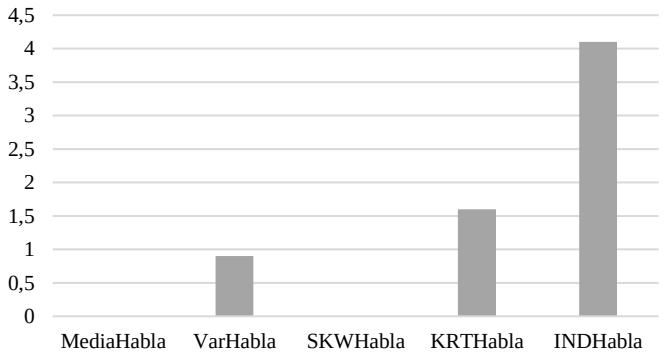


Figura 8-1 Selección de características: resultados gráficos de la función *fscnca* aplicada las medidas temporales.

Las mejores características temporales siguiendo esta metodología han sido VarHabla ($\sigma_{t_s}^2$), KRTHabla ($Kurt_{t_s}$) e INDHabla (Ind_{t_s}), siendo esta última la más relevante con una amplia diferencia.

De igual manera, en la Tabla 8-15 se muestran los resultados del proceso de selección de características de Temperatura Emocional. Se acompaña de una gráfica de los mismos (Figura 8-2) en la que se puede apreciar el peso sobre el conjunto de cada una de ellas.

Tabla 8-15 Selección de características: resultados de la selección de medidas de Temperatura Emocional.

Medida	Relevancia
te_d	0,9
$\overline{te_c}$	2,1
$\sigma^2_{\mathrm{te_c}}$	0,15
$\tilde{\mu}_{te_{c_3}}$	0
$Kurt_{te_c}$	0

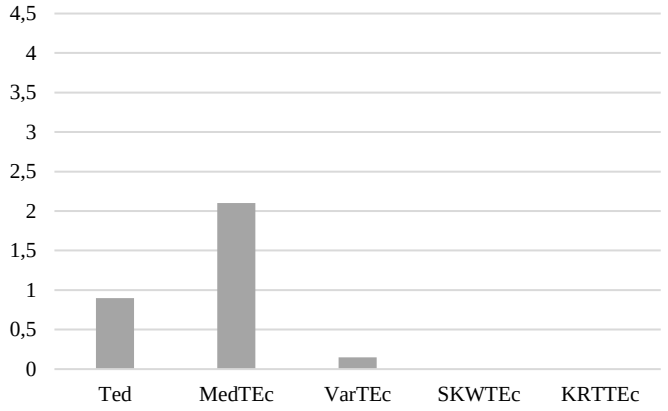


Figura 8-2 Selección de características: resultados gráficos de la función *fscnca* aplicada las medidas de Temperatura Emocional.

Respecto a las medidas de Temperatura Emocional, los resultados apuntan a que las mejores medidas son TEd (te_d), MediaTEc ($\overline{te_c}$), VarTEc ($\sigma_{te_c}^2$). Y entre ellas, la más relevante, MediaTEc ($\overline{te_c}$).

Conjunto óptimo

Dado que conocer las mejores características no implica saber cuál es el conjunto óptimo que permitiría obtener los mejores resultados de clasificación, en este apartado procedemos a estudiar los resultados de clasificación para cada una de las combinaciones posibles entre las mejores características obtenidas. El proceso seguido consiste en el uso de un modelo de clasificación de referencia que hemos considerado suficientemente potente para encontrar la combinación con mejores resultados: la Máquina de Vectores Soporte (SVM).

En la Tabla 8-16 se muestran los resultados de tasa de éxito [%], sensibilidad [%] y especificidad [%] obtenidos para cada una de las combinaciones de medidas temporales. Se ha realizado tanto para el conjunto de muestras del entrevistador automático como el humano y, también, para el conjunto completo de todas las muestras.

A partir de la Tabla 8-16 puede apreciarse que la combinación de medidas temporales que mejores resultados arroja es la misma, tanto si se usan las muestras del entrevistador automático, humano o el conjunto total de muestras: VarHabla ($\sigma_{t_s}^2$), KRTHabla ($Kurt_{t_s}$) e INDHabla (Ind_{t_s}). El conjunto óptimo de medidas temporales sería, por tanto, la combinación de estas tres variables.

Tabla 8-16 Conjunto óptimo: comparativa entre las diferentes combinaciones de las características temporales más relevantes (A: entrevistador automático, H: entrevistador humano, ACC: tasa de éxito de la clasificación, S: sensibilidad, E: especificidad), tipo de respuesta biclase, modelo SVM, validación cruzada 5 *folds*.

			Conjunto de características							
Entrev.	**Train/ Test**	**Clasif.**	$\bar{t}_S$	$\sigma^2_{t_S}$	$\tilde{\mu}_{t_{S_3}}$	$Kurt_{t_s}$	Ind_{t_s}	**ACC [%]**	**S [%]**	**E [%]**
A	**5 k-fold**	**SVM**					x	80,3	70	88
A	**5 k-fold**	**SVM**		x			x	80,7	69	90
A	**5 k-fold**	**SVM**		x		x	x	82,7	81	86
A	**5 k-fold**	**SVM**	x	x	x	x	x	80,7	69	90
H	**5 k-fold**	**SVM**					x	88,7	80	93
H	**5 k-fold**	**SVM**		x			x	88,7	80	93
H	**5 k-fold**	**SVM**		x		x	x	90,1	88	91
H	**5 k-fold**	**SVM**	x	x	x	x	x	90	84	93

En la misma línea que el análisis anterior, hemos buscado la combinación óptima para las medidas de Temperatura Emocional. Los resultados obtenidos se muestran en la Tabla 8-17.

Tabla 8-17 Conjunto óptimo: comparativa entre las diferentes combinaciones de las características de Temperatura Emocional más relevantes (A: entrevistador automático, H: entrevistador humano, ACC: tasa de éxito de la clasificación, S: sensibilidad, E: especificidad), tipo de respuesta biclase, modelo SVM, validación cruzada 5 *folds*.

			Conjunto de características							
Entrev.	**Train/ Test**	**Clasif.**	te_d	$\overline{te_c}$	$\sigma^2_{te_c}$	$\tilde{\mu}_{te_{c3}}$	$Kurt_{te_c}$	**ACC [%]**	**S [%]**	**E [%]**
A	**5 k-fold**	**SVM**		x				57,4	20	87
A	**5 k-fold**	**SVM**	x	x				57,4	20	87
A	**5 k-fold**	**SVM**	x	x	x			61,8	23	89
A	**5 k-fold**	**SVM**	x	x	x	x	x	60,2	20	93
H	**5 k-fold**	**SVM**		x				64,8	48	72
H	**5 k-fold**	**SVM**	x	x				64,8	48	72
H	**5 k-fold**	**SVM**	x	x	x			64,8	48	72
H	**5 k-fold**	**SVM**	x	x	x	x	x	64,8	48	72

En este caso, a tenor de los resultados mostrados en la Tabla 8-17, existen dos combinaciones posibles que podrían ofrecer los mejores valores de clasificación. La primera de ellas sería utilizando las medidas TEd (te_d), MediaTEd ($\overline{te_c}$) y VarTEd ($\sigma^2_{te_c}$) y la segunda el conjunto completo de todas las medidas de Temperatura Emocional. En cualquier caso, puede apreciarse que los resultados en términos de exactitud y sensibilidad son considerablemente inferiores que los obtenidos de las medidas temporales.

Por último, hemos probado realizando la clasificación con lo que, a partir de los resultados anteriores, hemos considerado la mejor combinación de medidas temporales y de Temperatura Emocional. Asimismo, se ha incluido también en la comparación los resultados de utilizar las diez características en conjunto. En la Tabla 8-18 se muestran estos resultados. Por su parte, las tres combinaciones con mejores resultados para cada tipo de muestra (habla inducida o espontánea) se recogen en la Tabla 8-19.

Tabla 8-18 Conjunto óptimo: comparativa entre las diferentes combinaciones del conjunto total de diez características (A: entrevistador automático, H: entrevistador humano, ACC: tasa de éxito de la clasificación, S: sensibilidad, E: especificidad), tipo de respuesta biclase, modelo: SVM, validación cruzada 5 *folds*.

			Conjunto de características												
Entrev.	**Train/ Test**	**Clasif.**	$\bar{t}_S$	$\sigma^2_{t_S}$	$\widetilde{\mu}_{t_{S_3}}$	$Kurt_{t_s}$	Ind_{t_s}	te_d	$\overline{te_c}$	$\sigma^2_{te_c}$	$\widetilde{\mu}_{te_{c_3}}$	$Kurt_{te}$	**ACC [%]**	**S [%]**	**E [%]**
A	5 k-fold	SVM	x	x	x	x	x	x	x	x	x	x	79,1	89	67
A	5 k-fold	SVM		x		x	x	x	x	x			80,3	70	88
A	5 k-fold	SVM		x		x	x	x	x				81,5	71	90
A	5 k-fold	SVM				x	x	x	x				79,9	71	87
A	5 k-fold	SVM		x		x	x		x				82,7	74	90
A	5 k-fold	SVM				x	x		x				80,3	68	90
H	5 k-fold	SVM	x	x	x	x	x	x	x	x	x	x	88,7	76	96
H	5 k-fold	SVM		x		x	x	x	x	x			88,7	84	91
H	5 k-fold	SVM		x		x	x	x	x				84,5	80	87
H	5 k-fold	SVM				x	x	x	x				87,3	80	91
H	5 k-fold	SVM		x		x	x		x				87,3	84	89
H	5 k-fold	SVM				x	x		x				87,3	84	89

Tabla 8-19 Conjunto óptimo: resumen de las tres mejores combinaciones de características para cada tipo de muestra (A: entrevistador automático, H: entrevistador humano, ACC: tasa de éxito de la clasificación, S: sensibilidad, E: especificidad, Coincid.: número de veces que la medida es incluida en una de las mejores combinaciones de características), tipo de respuesta biclase, modelo: SVM, validación cruzada 5 *folds*.

Entrev.	**Train/ Test**	**Clasif.**	$\bar{t}_S$	$\sigma^2_{t_S}$	$\widetilde{\mu}_{t_{S_3}}$	$Kurt_{t_s}$	Ind_{t_s}	te_d	$\overline{te_c}$	$\sigma^2_{te_c}$	$\widetilde{\mu}_{te_{c_3}}$	$Kurt_{te}$	**ACC [%]**	**S [%]**	**E [%]**
A	5 k-fold	SVM		x		x	x						82,7	81	86
A	5 k-fold	SVM						x	x	x			61,8	23	89
A	5 k-fold	SVM		x		x	x		x				82,7	74	90
H	5 k-fold	SVM		x		x	x						90,1	88	91
H	5 k-fold	SVM	x	x	x	x	x	x	x	x	x	x	88,7	76	96
H	5 k-fold	SVM		x		x	x	x	x	x			88,7	84	91
		Coincid.	1	5	1	5	5	3	4	3	1	1			

De la Tabla 8-19 se extrae que existen tres combinaciones aproximadamente comunes a los tres grupos de muestras. Sobre el conjunto total de diez características las que mejores resultados darían son:

- Conjunto Óptimo 1 (en adelante, CO1): VarHabla ($\sigma^2_{t_S}$), KRTHabla ($Kurt_{t_s}$) e INDHabla (Ind_{t_s}).
- Conjunto Óptimo 2 (en adelante, CO2): VarHabla ($\sigma^2_{t_S}$), KRTHabla ($Kurt_{t_s}$), INDHabla (Ind_{t_s}) y MediaTEc ($\overline{te_c}$).
- Conjunto Óptimo 3 (en adelante, CO3): VarHabla ($\sigma^2_{t_S}$), KRTHabla ($Kurt_{t_s}$), INDHabla (Ind_{t_s}), TEd (te_d), MediaTEc ($\overline{te_c}$) y VarTEc ($\sigma^2_{te_c}$).

A partir de estas medidas seleccionadas se desarrollará el siguiente apartado, a fin de encontrar los mejores resultados de clasificación.

8.2.2 Entrenamiento y prueba de modelos

En este apartado se presentan los resultados en términos de tasa de éxito [%], sensibilidad [%] y especificidad [%] obtenidos de aplicar las mejores combinaciones de características a seis modelos de clasificación optimizados: *tree*, *discriminant*, kNN, Naive Bayes, SVM y *ensemble*. Cabe mencionar que, para el método optimizado tipo *ensemble*, el *software* busca los mejores resultados ajustando los hiperparámetros de los modelos AdaBoost, RUSBoost, LogitBoost (clasificaciones binarias), GentleBoost (clasificaciones binarias) y Bag. Para estas simulaciones se ha analizado el caso de que la respuesta esté basada tanto en la variable Enfermedad (población sana/patológica) como en la variable Grados (población sana/leve/moderada) aplicando, en todos los casos, validación cruzada 5 *folds*.

Clasificación multivariante basada en presencia o ausencia de enfermedad

A continuación, en la Tabla 8-20 se muestran los resultados de clasificación para cada entrevistador aplicándoles cada uno de los conjuntos óptimos de características definidos, así como el conjunto total de diez medidas cuando la respuesta se basa en una clasificación biclase: sano/patológico. Para los dos entrevistadores encontramos que los mejores resultados en términos de tasa de éxito y sensibilidad (primando sensibilidad por encima de tasa de éxito) se concentran en torno a los clasificadores *ensemble*, para el entrevistador automático, y en torno al modelo SVM para el entrevistador humano.

Tabla 8-20 Clasificadores: resultados de clasificación para cada conjunto óptimo de características y para en conjunto total (ACC: tasa de éxito de la clasificación, S: sensibilidad, E: especificidad), tipo de respuesta biclase, modelo: varios clasificadores, validación cruzada 5 *folds*.

	Entrevistador automático			Entrevistador humano		
Modelo	**CONJUNTO**	**ACC [%]**	**S/E [%]**	**CONJUNTO**	**ACC [%]**	**S/E [%]**
Discriminant	CO1	79,5	69/88	CO1	85,9	80/89
Discriminant	CO2	80,7	70/89	CO3	85,9	80/89
Discriminant	TODAS	79,1	68/88	TODAS	84,5	76/89
Ensemble	CO1	82,3	77/87	CO1	87,3	80/91
Ensemble	CO2	81,1	75/86	CO3	85,9	76/91
Ensemble	TODAS	80,3	70/88	TODAS	88,7	80/93
kNN	CO1	80,3	71/88	CO1	87,3	72/96
kNN	CO2	78,3	65/89	CO3	83,1	76/87
kNN	TODAS	78,3	61/92	TODAS	85,9	68/96
Naive-Bayes	CO1	78,3	67/88	CO1	83,1	72/89
Naive-Bayes	CO2	77,1	68/85	CO3	81,7	76/85
Naive-Bayes	TODAS	75,9	68/83	TODAS	78,9	72/83
SVM	CO1	82,3	71/91	CO1	91,5	84/96
SVM	CO2	81,5	71/90	CO3	88,7	84/91
SVM	TODAS	81,1	68/91	TODAS	88,7	84/91
Tree	CO1	81,1	70/90	CO1	88,7	80/93
Tree	CO2	81,9	68/93	CO3	84,5	76/89
Tree	TODAS	80,3	65/93	TODAS	88,7	84/91

En la Tabla 8-21 y Figura 8-3 se muestra el mejor modelo para el entrevistador automático y la configuración optimizada que se ha aplicado para obtenerlo.

Tabla 8-21 Clasificadores: selección de los mejores clasificadores (entrevistador: automático, clasificación: enfermedad).

Clasificador	Conjunto óptimo	ACC [%]	Sensibilidad [%]	Especificidad [%]
Ensemble (AdaBoost)	**CO1**	82,3	74	90

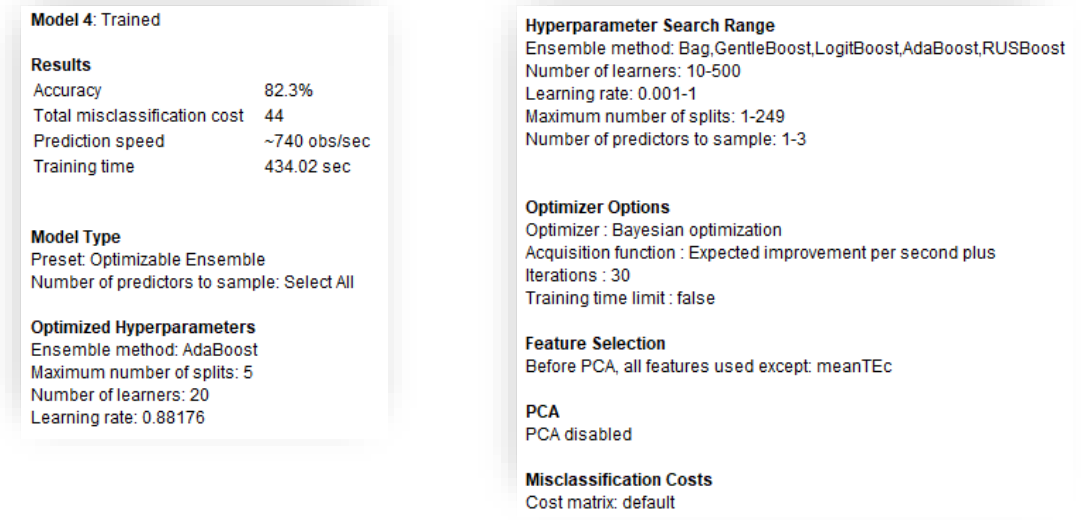


Figura 8-3 Clasificadores: informe generado para el modelo de entrenamiento Ensemble_CO1: ajuste de hiperparámetros y configuración optimizada (entrevistador: automático, clasificación: enfermedad).

Por su parte, en la Tabla 8-22 y Figura 8-4 se muestra el mejor modelo para el entrevistador humano y la configuración optimizada que se ha aplicado para obtenerlo.

Tabla 8-22 Clasificadores: selección de mejores clasificadores (entrevistador: humano, clasificación: enfermedad).

Clasificador	Conjunto óptimo	ACC [%]	Sensibilidad [%]	Especificidad [%]
SVM	**CO1**	91,5	84	96

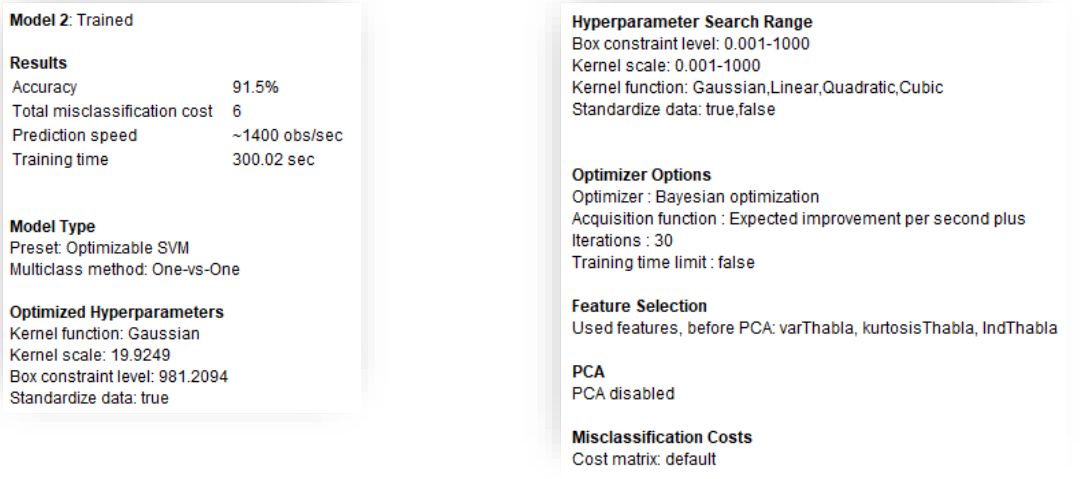


Figura 8-4 Clasificadores: informe generado para el modelo de entrenamiento SVM_CO1: ajuste de hiperparámetros y configuración optimizada (entrevistador: humano, clasificación: enfermedad).

Clasificación multivariante basada en diferentes grados de enfermedad

De la misma manera que en el apartado previo, en la Tabla 8-23 se muestran los resultados de clasificación para cada entrevistador aplicándoles cada uno de los conjuntos óptimos de características (así como el conjunto total de características) cuando la respuesta se basa en tres clases o grados: sano, leve y moderado.

Tabla 8-23 Clasificadores: resultados de clasificación para cada conjunto óptimo de características y para en conjunto total (ACC: tasa de éxito de la clasificación, S: sensibilidad, E: especificidad), tipo de respuesta basada en grados, modelo: varios clasificadores, validación cruzada 5 *folds*.

Modelo	Entrevistador automático				Entrevistador humano			
	Conjunto óptimo	ACC [%]	S [%]	E [%]	Conjunto óptimo	ACC [%]	S [%]	E [%]
Discriminant	CO1	70,7	64,86	92,03	CO1	81,7	76,00	91,30
Discriminant	CO2	69,5	63,96	91,30	CO3	77,5	72,00	86,96
Discriminant	Todas	67,5	63,06	89,13	Todas	78,9	80,00	91,30
Ensemble	CO1	75,1	70,27	92,75	CO1	80,3	76,00	89,13
Ensemble	CO2	74,3	69,37	93,48	CO3	78,9	72,00	89,13
Ensemble	Todas	72,7	66,67	92,75	Todas	80,3	76,00	89,13
kNN	CO1	69,1	63,06	93,48	CO1	84,5	76,00	95,65
kNN	CO2	69,5	63,06	93,48	CO3	80,3	68,00	93,48
kNN	Todas	63,1	55,86	88,41	Todas	80,3	68,00	91,30
Naive-Bayes	CO1	67,9	65,77	86,23	CO1	77,5	72,00	89,13
Naive-Bayes	CO2	68,7	63,06	89,13	CO3	78,9	72,00	91,30
Naive-Bayes	Todas	63,9	63,96	85,51	Todas	74,7	80,00	78,26
SVM	CO1	72,7	66,67	94,93	CO1	84,5	84,00	91,30
SVM	CO2	73,5	67,57	94,93	CO3	81,7	72,00	91,30
SVM	Todas	68,3	66,67	87,68	Todas	77,5	76,00	91,30
Tree	CO1	72,3	68,47	91,30	CO1	81,7	80,00	89,13
Tree	CO2	71,9	67,57	91,30	CO3	81,7	80,00	89,13
Tree	Todas	72,7	66,67	92,75	Todas	81,7	80,00	89,13

Para el entrevistador automático el clasificador con mejores valores de sensibilidad sería el modelo *ensemble* para el conjunto óptimo CO1. Para el entrevistador humano el clasificador que pasaría a tener los mejores resultados es SVM, también para el conjunto de características CO1.

A continuación, en la Tabla 8-24 y Figura 8-5 se muestra el mejor modelo para el entrevistador automático y la configuración optimizada que se ha aplicado para obtenerlo, respectivamente.

Tabla 8-24 Clasificadores: selección de los mejores clasificadores (entrevistador: automático, clasificación: grados).

Clasificador	Conjunto óptimo	ACC [%]	Sensibilidad [%]	Especificidad [%]
Ensemble (Bag)	**CO1**	73,90	71,17	91,30

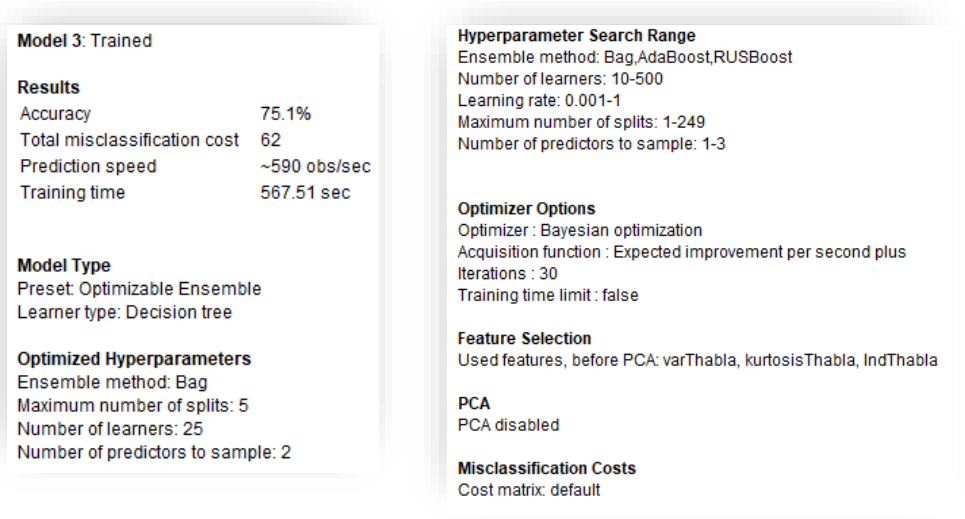


Figura 8-5 Clasificadores: informe generado para el modelo de entrenamiento Ensemble_CO1: ajuste de hiperparámetros y configuración optimizada (entrevistador: automático, clasificación: grados).

Por su parte, en la Tabla 8-25 y Figura 8-6 se muestra el mejor modelo para el entrevistador humano y la configuración optimizada que se ha aplicado para obtenerlo, respectivamente.

Tabla 8-25 Clasificadores: selección de mejores clasificadores (entrevistador: humano, clasificación: grados).

Clasificador	Conjunto óptimo	ACC [%]	Sensibilidad [%]	Especificidad [%]
SVM	**CO1**	84,5	84,00	91,30

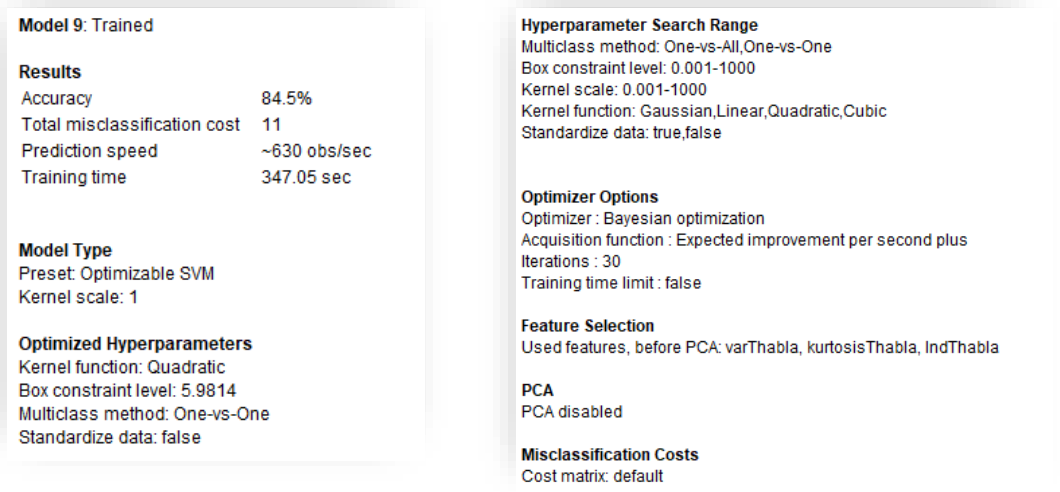


Figura 8-6 Clasificadores: informe generado para el modelo de entrenamiento SVM_CO1: ajuste de hiperparámetros y configuración optimizada (entrevistador: humano, clasificación: grados).

Capítulo 9 Análisis de los resultados

9.1 Análisis del estado del arte

Debido al envejecimiento progresivo de la población, en las últimas décadas, ha habido un interés creciente y un aumento significativo en el número de recursos invertidos y publicaciones escritas sobre enfermedades como la enfermedad de Alzheimer (AD). La AD es uno de los grandes desafíos de nuestra sociedad y, actualmente, las líneas de investigación desarrolladas son muy diversas. En este estudio, se ha llevado a cabo una búsqueda sistemática para relacionar las diferentes técnicas de procesamiento de voz y habla aplicadas a la detección de la AD. La evolución o tendencia respecto al número de publicaciones centradas en el estudio de la demencia en general y aquéllas basadas en la AD y el procesamiento automático de voz en particular, se representa en la Figura 9-1.

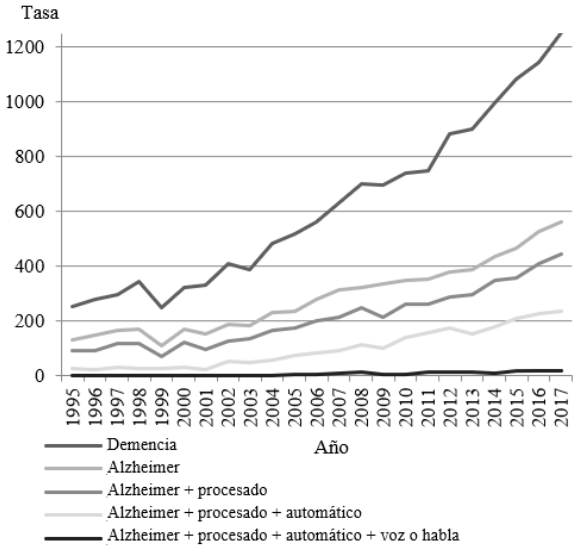


Figura 9-1 Publicaciones identificadas sobre demencia frente a aquellas centradas en la aplicación del procesado automático de voz para la detección o control evolutivo de la AD, desde 1995 [36].

Con el fin de analizar la tendencia de los estudios centrados en la detección de la AD mediante el análisis automático de voz, hemos comenzado dividiendo

cuidadosamente las publicaciones localizadas según el tipo de características utilizadas: convencionales, sin duda las más utilizadas y extensas, y las no convencionales, menos conocidas y aún por explorar. Una tercera parte aquí desarrollada se refiere a las técnicas de aprendizaje profundo, *Deep Learning*. Si bien no encajan exactamente en nuestra clasificación anterior, parece relevante incluirlas en el análisis ya que ofrecen soluciones interesantes para detectar adecuadamente la AD y, en un futuro no muy lejano, distinguirla de otros impedimentos, demencias o patologías.

Es importante aclarar que la presente tesis no ha entrado en definir factores clínicos ni en una clasificación detallada de estas poblaciones. En este sentido, por ejemplo, dos poblaciones etiquetadas como MCI podrían tener diferentes discapacidades, debido a subtipos o progresiones y, por lo tanto, diferencias parciales en sus patrones de voz. Dentro del alcance de la revisión realizada, no es posible explorar más en profundidad dichas implicaciones específicas que escapan a los objetivos de esta tesis. La búsqueda de una subdivisión más detallada es claramente una tarea muy interesante y necesaria, donde no sólo se pueden clasificar diferentes tipos de MCI, sino también diferentes patologías y sus correspondientes patrones particulares de voz.

A partir de la revisión crítica del estado del arte podemos decir que, desde que surgieran las primeras publicaciones en el campo, aproximadamente en 2005, el 78% de las investigaciones se han centrado en el uso de parámetros convencionales, principalmente en la duración de los segmentos sonoros y no sonoros, tono, amplitud y periodicidad, así como otras características obtenidas del análisis temporal, frecuencial y cepstral. Estas variables, como hemos visto, han proporcionado sin duda datos valiosos sobre procesos cognitivos y sus valores se han relacionado directamente con el estado específico de la AD. Cada autor ha realizado diferentes interpretaciones de estos parámetros y, en ese sentido, ha relacionado los déficits cognitivos con fenómenos comunicativos y características como la prosodia, la fonación, la articulación y las características vocales y paralingüísticas, entre otras.

Del mismo modo, a través de combinaciones de varios parámetros, se han definido diferentes conceptos como la calidad de la voz o el habla, o la temperatura emocional (TE). Estos conceptos, aplicados correctamente, han demostrado ser buenos indicadores de la AD. Técnicas más elaboradas, como el análisis automático del habla espontánea (ASSA), se presentan como métodos que involucran diferentes atributos combinados de la voz (ASSA, por ejemplo, duraciones, energía a corto plazo y *spectral centroid*), y han demostrado ofrecer información extremadamente relevante.

Dentro del campo del Aprendizaje Automático o *Machine Learning*, si hay algo a lo que apunta la mayoría de estudios, es a la importancia del número de medidas obtenidas en el proceso de extracción de las mismas (*feature extration*). Diferentes estudios ya han trabajado en este aspecto y, a la luz de los resultados, está claro que un conjunto más reducido de características conduce a mejores resultados en términos, por ejemplo, de exactitud. Esto es especialmente importante cuando el número de muestras es relativamente reducido: si tenemos menos características, tendremos una mayor capacidad de generalización, así como un menor coste computacional. Por tanto, en estos casos, una fase de selección de características (*feature selection*) es necesaria y deseable antes de llevar a cabo un proceso de clasificación.

En cuanto a las tareas verbales realizadas por los sujetos en los diferentes estudios, puede concluirse que han consistido principalmente en una serie de muestras de habla espontánea. Según nuestra investigación, ésta es la tarea más extendida y, tal vez, la que proporciona los resultados más relevantes debido al rendimiento claramente deficiente que los pacientes con AD presentan a la hora de comunicarse. Aunque menos frecuente, otros estudios han desarrollado tareas de lectura, repetición o recuento, entre otros (Figura 9-2).

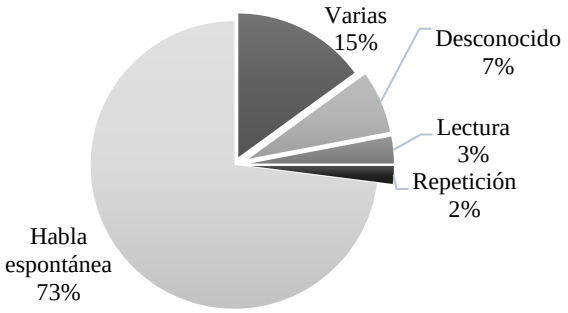


Figura 9-2 Tareas verbales más habituales para el análisis de voz o habla aplicado a la detección de AD. Basado en las publicaciones localizadas [36].

Las tareas desarrolladas por los locutores merecen especial atención ya que las características que pueden medir el deterioro en un grupo de pacientes y su significado varían ampliamente según la actividad que realice el sujeto. Sería deseable, por tanto, un análisis más profundo, por ejemplo, para esclarecer cuál es el significado de una pausa realizada durante una tarea de fluidez verbal, ya que esto puede ser indicativo de problemas de memoria semántica pero no es necesariamente cierto aplicado a pausas realizadas durante una tarea de conteo. Los estudios localizados en esta revisión no han sido clasificados teniendo en cuenta este hecho y, por tanto, en ese sentido, sería interesante profundizar más sobre la cuestión. Por su parte, como denominador común en los trabajos analizados tenemos que, una vez se realiza el proceso de extracción de características, a partir de una tarea verbal concreta, se procede al análisis y/o la clasificación de los datos. Tal y como hemos podido analizar, en la mayoría de los casos, esto se lleva a cabo a través de modelos clasificadores SVM, kNN, RF o Naive Bayes, en menor medida, y clasificadores LDA (Figura 9-3).

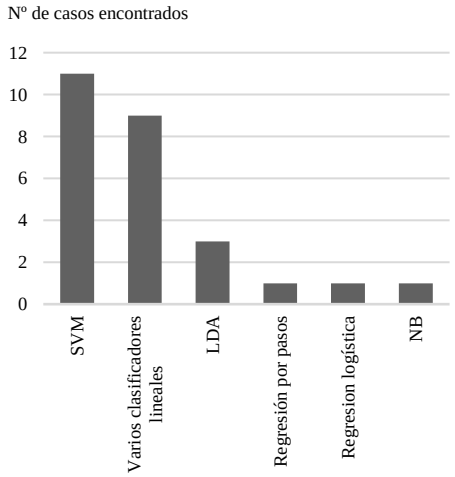


Figura 9-3 Clasificadores más populares empleados en el análisis de voz aplicado a la detección de AD. Basado en las publicaciones localizadas [36].

Los procesos de extracción y selección de las diferentes características, así como el uso de los clasificadores mencionados anteriormente, en general, han demostrado tener buenos resultados en cuanto a la evaluación objetiva del estado de la AD. Además, ofrecen la posibilidad de ser aplicados en un futuro a otros trastornos neurodegenerativos como puede ser la enfermedad de Parkinson.

Sin embargo, a pesar de este hecho y de sus buenos resultados, existen otros aspectos a tener en cuenta en el análisis del habla aplicada a la detección de la AD. Está claro que se pueden encontrar, por ejemplo, diferencias lingüísticas, no sólo entre hablantes de diferentes idiomas sino también entre sus propios dialectos. Este hecho podría ser un problema con respecto a la implementación de una herramienta a gran escala. No todos nos expresamos de la misma manera, hablamos a la misma velocidad o usamos la misma cantidad de palabras para comunicarnos. En este punto, es interesante contemplar si sería posible crear una herramienta válida para ayudar al diagnóstico a nivel global. Algunos investigadores habrían encontrado una respuesta dentro del análisis emocional. Como se ha demostrado, las diferencias de los estados emocionales pueden considerarse como uno de los criterios de evaluación más importantes para medir el rendimiento cognitivo. Dado que las emociones son características humanas universales e intrínsecas, reconocerlas en el proceso comunicativo ha causado un gran interés en la comunidad científica.

En esta línea, varios estudios han incluido características clásicas como el tono, la intensidad y variación de componentes de frecuencia y, más recientemente, la Temperatura Emocional (una combinación de características prosódicas y paralingüísticas) en el análisis emocional. Estos estudios aplican análisis de la respuesta emocional, ERA, que utiliza diferentes características lineales y, combinados con medidas de TE y habla espontánea, han conseguido una buena discriminación entre pacientes con AD y sujetos HC, alcanzando un 97% de exactitud. Del mismo modo, se han realizado varios estudios que combinan ASSA con estas otras técnicas y también ofrecen resultados prometedores. Estos métodos se basan en características acústicas y estacionarias de la señal de voz. La mayoría de ellos utilizan métodos estacionarios en el dominio de la frecuencia, como el espectro de potencia de Fourier.

Por su parte, también se han publicado estudios que utilizan transcripciones y la aplicación de detectores de actividad de voz, especialmente interesantes ya que, además del análisis acústico de la señal de voz o habla, también ofrecen análisis léxicos, semánticos, de puntuación y análisis sintácticos del proceso de comunicación.

En los últimos años, los aspectos no lineales y no estacionarios relacionados con los cambios dinámicos en la señal del habla afectados por el deterioro cognitivo parecen haber despertado también un gran interés. Desde el año 2012 aproximadamente, cada vez más estudios en el campo subrayan la necesidad de trabajar con características no convencionales. Actualmente, representan aproximadamente el 22% de las publicaciones más relevantes. Por su parte, algunos investigadores han propuesto que los cambios cognitivos sutiles en los estadios tempranos y preclínicos podrían detectarse mediante fractales. Por esta razón, haciendo uso concretamente de la Dimensión Fractal de Higuchi, técnicas como ASSA y ERA ya se han incluido en el conjunto de características acústicas y estacionarias y también con clasificadores muy potentes como

son los basados en redes neuronales MLP. Los resultados obtenidos mediante estos métodos han resultado prometedores e incrementan aún más las expectativas que utilizando exclusivamente características lineales convencionales.

Otras investigaciones también recientes han demostrado que no es viable caracterizar con una sola relación de escala aquellas geometrías y eventos que evolucionan naturalmente. Estos trabajos sostienen que un sistema como la voz se caracteriza mejor por medio de multifractales. Entre estos métodos de análisis, el MFDFA se ha aplicado con éxito al análisis del habla y sus resultados son más fiables en comparación con otros métodos como el Análisis Wavelet, la Transformación Wavelet Discreta o el Módulo Máximo de Transformada Wavelet, entre otros. Además, estos mismos sistemas también se han aplicado para analizar eventos como la dinámica de la frecuencia cardíaca, el aumento de las neuronas o las series económicas de tiempo.

En cualquier caso, el crecimiento de la Inteligencia Artificial y, específicamente, del Aprendizaje Automático, *Machine Learning* en inglés, ha revolucionado las perspectivas generales en el análisis de voz y habla al dotarlos de sistemas más complejos. El Aprendizaje Automático ofrece análisis estadísticos y matemáticos y una multitud de técnicas de extracción de características, así como de clasificación automática.

Por su parte, el Aprendizaje Profundo o *Deep Learning* da un paso más allá respecto a las técnicas de *Machine Learning* al disponer y trabajar con conjuntos mucho de mayores de datos y técnicas de clasificación más sofisticadas como son las basadas en capas de aprendizaje. En cualquier caso, actualmente, sigue siendo un campo bajo estudio y desarrollo. La Figura 9-4 representa un gráfico comparativo donde se puede observar, en concreto, la tendencia creciente en el número de artículos publicados en torno al uso de técnicas de Aprendizaje Profundo y redes neuronales aplicadas a la detección de AD a partir de la señal de voz.

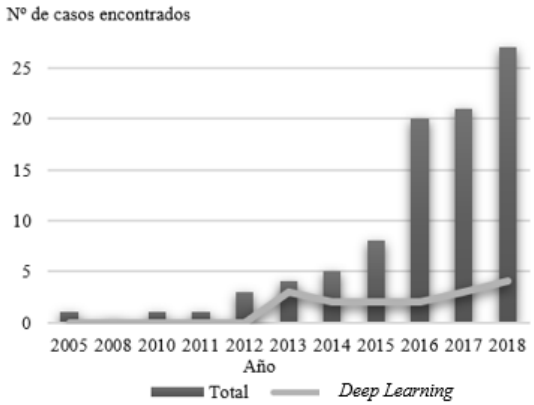


Figura 9-4 Artículos publicados sobre aprendizaje profundo aplicado a la detección automática del procesamiento de voz y habla. Basado en las publicaciones localizadas [36].

El aumento progresivo en el número de publicaciones se habría producido principalmente desde el año 2012, donde los modelos más utilizados a este respecto han sido: CNN, RNN y GCNN, aplicados tanto a características lingüísticas (convencionales y no convencionales) como paralingüísticas.

A día de hoy, las innovadoras técnicas de Aprendizaje Profundo parecen ofrecer soluciones interesantes en la detección precoz y control evolutivo de la AD. Habrían colocado el procesado de señales entre los métodos más prometedores dotando de soluciones a sistemas tan complejos como es el proceso comunicativo, la producción de voz y habla y los procesos cognitivos humanos.

9.2 Análisis de las bases de datos

Las bases de datos son un aspecto fundamental de cualquier investigación ya que, sólo a partir de ellas, podemos desarrollar nuestros estudios y análisis experimentales. En esta tesis una de las tareas que hemos llevado a cabo ha sido la de localizar y clasificar una serie de bases de datos relacionadas con el uso del lenguaje en pacientes con AD con el fin de comprender mejor su papel y situación actual en el campo. A partir de una revisión exhaustiva del estado del arte, hemos encontrado, además de una gran escasez en el número de bases de datos localizadas, una gran diversidad en cuanto a cómo se realizan las grabaciones de los sujetos. Divergencias, por ejemplo, respecto a los aspectos relacionados con el proceso de grabación, con la automatización de las entrevistas o con las tareas lingüísticas realizadas por los sujetos, que difieren mucho de un estudio a otro. Aunque aspectos como los mencionados anteriormente influyen en las grabaciones, no son los únicos; hay muchas otras variables como el idioma, el entorno o simplemente los métodos de preprocesamiento utilizados que hacen que cada base de datos se conforme en diferentes condiciones.

Sin embargo, cabe mencionar que existe cierta inclinación en cuanto al tipo de tarea lingüística realizada: hasta un 80% de los casos analizados utilizan el habla espontánea en sus grabaciones, entendiendo el habla espontánea como aquellas tareas en las que se plantean preguntas al sujeto y se le da un tiempo limitado y relativamente largo para expresarse libremente. Sin embargo, no tenemos pruebas claras de si otras tareas lingüísticas, como la lectura, podrían aportarnos más o menos información o, incluso, ser complementarias entre sí. Por otro lado, hemos podido observar que tan sólo el 18% de las bases de datos localizadas corresponden a estudios longitudinales. Este tipo de estudio es de especial interés porque permite el análisis de la variable tiempo sobre las muestras, lo que sin duda es un reflejo del progresivo deterioro del lenguaje que sufren estos pacientes. Entre las bases de datos localizadas, cabe destacar que sólo una de ellas automatizó el proceso de entrevista con el sujeto mediante avatares computarizados. Sin embargo, no sabemos cómo este hecho pudo afectar efectivamente sobre los resultados finales. Conocer el potencial de estas herramientas y el que podrían alcanzar a medida que se fueran desarrollando y perfeccionando se presenta, a día de hoy, como tarea obligada en el campo del análisis automático de voz para la detección precoz de la AD.

Para conocer cómo afecta el proceso de automatización de entrevistas a la toma de muestras y a los resultados de los análisis estadísticos posteriores, en esta tesis se ha analizado la base de datos Cross-Sectional Alzheimer Prognosis R2019 (CSAP-R19). Esta base de datos consta de dos tipos de grabaciones: habla espontánea (entrevistador

humano) y habla inducida (entrevistador automático). Las muestras de habla inducida se han recogido mediante el *software* Prognosis perteneciente al Proyecto Prognosis de la Universidad de Las Palmas de Gran Canaria, en el que se incluye esta tesis, y que se viene desarrollando desde el año 2014. En este sentido, cabe destacar dos ventajas principales de la base de datos utilizada. En primer lugar, se trata de una base de datos en la que se recogen dos claras metodologías para grabar la voz (tenemos un mismo sujeto que participa en ambos tipos de grabaciones) y cabe mencionar que hasta el momento no hemos encontrado trabajos previos similares. La segunda ventaja reside en el potencial que tendrían estos registros a partir de los cuales se podría, llegado el caso, validar las metodologías planteadas y la automática en particular.

Por su parte, las ventajas que se han podido identificar en cuanto a la recogida automática de muestras han sido numerosas a pesar de que, a día de hoy, siguen siendo escasos los sistemas que lo aplican. Entre ellas su escalabilidad, rapidez, objetividad o bajo coste son sólo algunas de esas ventajas que ofrecería. La escasez en cuanto a los sistemas que sí lo han aplicado es uno de los principales motivos por los que todavía no se ha podido demostrar hasta qué punto para un mismo sujeto los resultados obtenidos de un proceso de entrevistas automático son realmente útiles.

9.3 Análisis estadístico: medidas temporales

A fin de conocer más sobre los entrevistadores automáticos y su posible aplicación en la detección precoz de AD a partir de la voz, en esta tesis hemos comparado diferentes grabaciones obtenidas de un entrevistador humano y uno automático.

Para ello hemos realizado un proceso de extracción de características temporales basadas en la duración de las fonaciones de los sujetos HC y pacientes AD para los dos tipos de discurso contenidos en la base de datos CSAP-R19. Para comprender su comportamiento y dar sentido a los valores obtenidos durante el proceso de extracción de características hemos aplicado diferentes análisis univariantes y multivariantes así como diferentes modelos de clasificación teniendo en cuenta las tres poblaciones contempladas en CSAP-R19: sujetos de control (HC), AD leve (AD1) y AD moderado (AD2).

9.3.1 Análisis univariante

En primer lugar, hemos estudiado por separado cada una de las cinco variables temporales definidas: Media del tiempo de habla ($\overline{t_s}$), Varianza del tiempo de habla ($\sigma_{t_s}^2$), *Skewness* del tiempo de habla ($\tilde{\mu}_{t_{s_3}}$), Curtosis del tiempo de habla ($Kurt_{t_s}$) e Índice del tiempo de habla (Ind_{t_s}).

El primer estudio realizado basado en el análisis de los anteriores estadísticos descriptivos refleja una tendencia clara en las duraciones de habla de los pacientes AD, sea cual sea el entrevistador utilizado. Concretamente, se puede observar a partir de la Tabla 7-1, cómo los valores de las variables MediaHabla ($\overline{t_s}$), VarHabla ($\sigma_{t_s}^2$) e INDHabla (Ind_{t_s}) disminuyen a medida que el grado de severidad de la AD aumenta.

En la Figura 9-5, Figura 9-6, Figura 9-7, Figura 9-8 y Figura 9-9 se puede cómo varían los valores medios de las cinco variables para cada entrevistador en función del grado de la enfermedad.

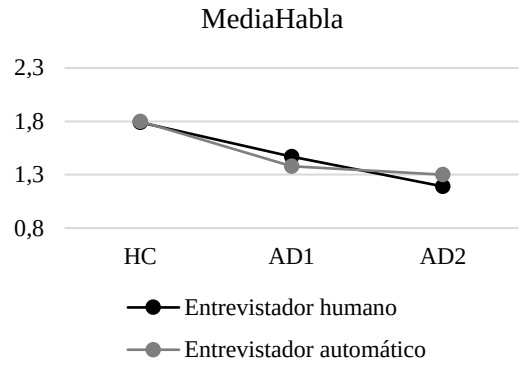


Figura 9-5 Estadísticos descriptivos: comparación de valores de MediaHabla ($\bar{t_S}$) para entrevistador humano y automático.

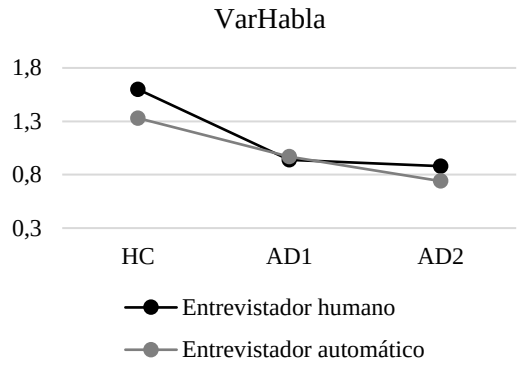


Figura 9-6 Estadísticos descriptivos: comparación de valores de VarHabla $(\sigma_{t_S}^2)$ para entrevistador humano y automático.

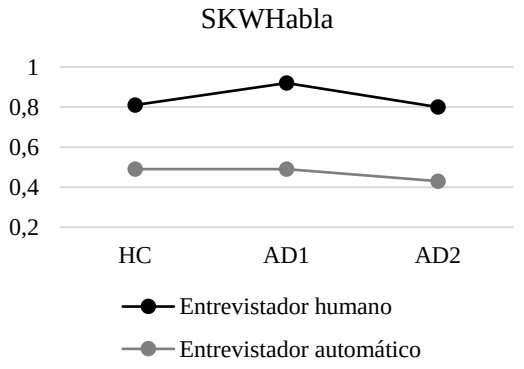


Figura 9-7 Estadísticos descriptivos: comparación de los valores de SKWHabla ($\tilde{\mu}_{t_{S_3}}$) para entrevistador humano y automático.

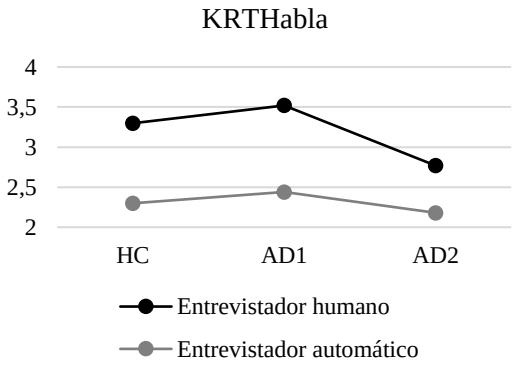


Figura 9-8 Estadísticos descriptivos: comparación de valores de KRTHabla ($Kurt_{t_s}$)para entrevistador humano y automático.

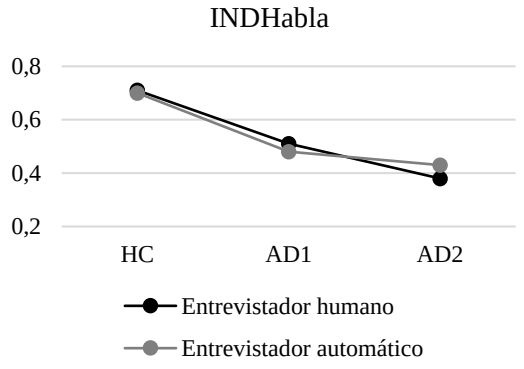


Figura 9-9 Estadísticos descriptivos: comparación de los valores de INDHabla (Ind_{t_s}) para entrevistador humano y automático.

Tanto las variables MediaHabla ($\bar{t_S}$) como VarHabla $(\sigma_{t_S}^2)$ e INDHabla (Ind_{t_S}) son mayores en los sujetos de control y los pacientes AD leves que en los pacientes AD

moderados, quienes tienen afectadas las capacidades comunicativas en mayor medida. Por su parte, a tenor de los valores mostrados en la Tabla 7-1, las variables SKWHabla ($\tilde{\mu}_{t_{S_3}}$) y KRTHabla ($Kurt_{t_s}$) no han resultado tan concluyentes al respecto.

En base a la tendencia que siguen las variables MediaHabla ($\bar{t_S}$), VarHabla ($\sigma^2_{t_S}$) e INDHabla (Ind_{t_s}), si bien es cierto que los estadísticos varían ligeramente en función del entrevistador utilizado, sí se puede demostrar que, en grabaciones con la misma duración, ambas metodologías siguen una tendencia similar reflejando mayores tiempos de habla en los sujetos sanos que en los patológicos. En este sentido, la idea planteada inicialmente en la que se proponía distinguir entre sujetos sanos y patológicos independientemente del método de grabación utilizado sí se cumple y, por tanto, se avalaría el planteamiento de que con ambos entrevistadores existen posibilidades de discriminar AD. Los valores obtenidos mediante un entrevistador u otro, para cada población estudiada y para cada variable, resultan ser muy similares.

Por su parte, a partir de los tres análisis no paramétricos realizados: Test de Wilcoxon, Test de Kruskal-Wallis y Test de la Mediana, hemos podido profundizar sobre cuáles de estas cinco variables temporales resultarían ser discriminante AD realizando cuatro comparaciones diferentes entre las poblaciones.

A partir de la Tabla 7-3, las variables que mejores resultados reportan son MediaHabla ($\bar{t_S}$) e INDHabla (Ind_{t_s}). En el caso de la variable VarHabla ($\sigma^2_{t_S}$), vemos que los valores son ligeramente peores, aunque seguiría la línea de las anteriores. Las variables SKWHabla ($\tilde{\mu}_{t_{S_3}}$) y KRTHabla ($Kurt_{t_s}$) en ninguno de los cuatro escenarios no paramétricos estudiados resultan ser discriminantes y, de nuevo, no se consideran concluyentes. De manera general para todas las variables, se puede observar que los peores resultados se obtienen en la comparación entre las poblaciones AD1-AD2 y HC-AD2. Probablemente esto se deba a que las poblaciones AD1 y AD2 constituyen conjuntos de muestras no lo suficientemente grandes y, además, muy similares entre sí. Concretamente el menor número de muestras de la base de datos es para la población AD2 (AD moderado), del que existen 15 pacientes grabados. En el caso de la población AD1 (AD leve), se han grabado 26 pacientes. Los sujetos de control entrevistados ascienden a 46. Podría ser necesario, por tanto, aumentar el número de muestras especialmente en el grupo AD2 para tener resultados más concluyentes al respecto.

En cualquier caso, en la comparación AD1-AD2 ninguna de las variables bajo estudio resultaría ser discriminante, al menos, tal y como se ha planteado este estudio. En la comparación realizada entre las poblaciones HC-AD2 la variable MediaHabla ($\bar{t_S}$) empeora sus resultados dejando de ser discriminante para el entrevistador humano, si bien es cierto que se encontraría justo al límite del valor fijado para *Prob*|*z*|, justo en el 5%. Los mejores resultados se obtienen para las variables MediaHabla ($\bar{t_S}$) e INDHabla (Ind_{t_s}), las cuales sí confirmarían su validez clara para discriminar entre las poblaciones HC-AD y HC-AD1.

9.3.2 Análisis multivariante

El análisis multivariante realizado ha sido el test paramétrico MANOVA. Los resultados recogidos en la Tabla 7-4 muestran cómo, para todos los estadísticos

analizados, el set de características utilizado es discriminante AD. Sin embargo, cuando las variables de agrupación son las poblaciones AD1 y AD2, para ambos entrevistadores los valores de los estadísticos están muy por encima del límite establecido para el *p-value* (0,05). Éste, en cualquier caso, no es un hecho aislado e iría en línea con los resultados obtenidos en las pruebas anteriores, como el estudio no paramétrico. En los tres test univariantes no paramétricos llevados a cabo, ninguna de las cinco variables por separado resultaba ser discriminante cuando lo que comparábamos eran las poblaciones AD1 y AD2. El motivo podría ser el expuesto anteriormente sobre el número de muestras disponibles para cada población a lo que se sumaría una mayor dificultad a la hora de diferenciar dos voces patológicas de diferentes grados.

Como puede deducirse a partir de los datos obtenidos, las muestras que ofrecen mejores resultados en cuanto a discriminación AD son las de habla espontánea, aquellas obtenidas con el entrevistador humano. Sin embargo, al llevar a cabo tanto en el análisis univariante como en el multivariante, hemos observado cómo la tendencia del entrevistador automático se asemeja considerablemente a la del humano. Hemos comprobado cómo en algunos casos, por ejemplo en el análisis univariante, las variables MediaHabla ($\bar{t}_S$), VarHabla ($\sigma_{t_S}^2$) e INDHabla (Ind_{t_s}) se acercan especialmente a los resultados del entrevistador humano para las tres poblaciones.

9.4 Análisis estadístico: Temperatura Emocional

9.4.1 Análisis univariante

El análisis univariante de las medidas de Temperatura Emocional se ha llevado a cabo siguiendo la misma estructura que se ha seguido con las medidas temporales. En primer lugar, hemos estudiado por separado cada una de las cinco variables definidas: Temperatura Emocional discreta (te_d), Media de la Temperatura Emocional continua ($\overline{te_c}$), Varianza de la Temperatura Emocional continua ($\sigma_{\mathrm{te_c}}^2$), *Skewness* de la Temperatura Emocional continua ($\tilde{\mu}_{te_{c_3}}$) y Curtosis de la Temperatura Emocional continua ($Kurt_{te_c}$).

El primer estudio realizado basado en el análisis de los estadísticos descriptivos refleja valores y tendencias similares en las variables TEd (te_d), SKWTEc ($\tilde{\mu}_{te_{c_3}}$) y KRTTEc ($Kurt_{te_c}$) con los dos entrevistadores. Para todas las características los valores obtenidos para los sujetos HC comparten un mismo punto de partida sea cual sea el entrevistador utilizado. Por su parte, la variable TEd (te_d) reflejaría la caída de los valores medios de Temperatura Emocional en los pacientes AD respecto a los sujetos HC en el caso del entrevistador humano. Los valores de KRTTEc ($Kurt_{te_c}$) denotan un menor número de datos concentrados en torno al valor medio TE (valores de KRTTEc ($Kurt_{te_c}$) menores de 3). Este hecho se acusa más para los pacientes de grado moderado además de repetirse para ambos entrevistadores. Respecto al resto de variables, a simple vista no se identifica un patrón o tendencia clara y, posiblemente, sea necesario otro tipo de análisis más profundo para detectarlo. En la Figura 9-10, Figura

9-11, Figura 9-12, Figura 9-13 y Figura 9-14 se representan los valores obtenidos para cada variable, grado y entrevistador.

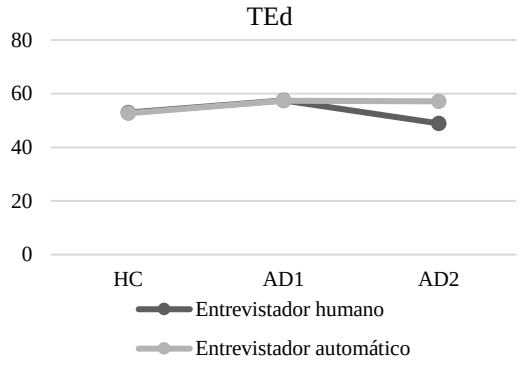


Figura 9-10 Estadísticos descriptivos: comparación de valores de TEd (te_d) para entrevistador humano y automático.

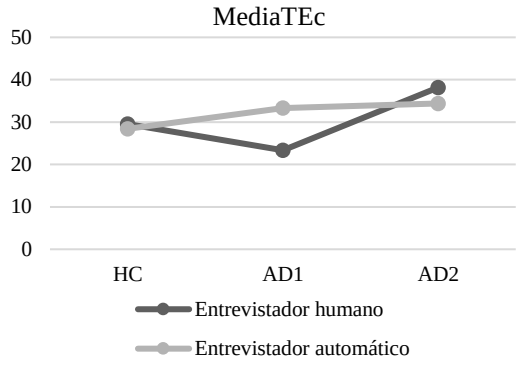


Figura 9-11 Estadísticos descriptivos: comparación de valores de MediaTEc ($\overline{te_c}$) para entrevistador humano y automático.

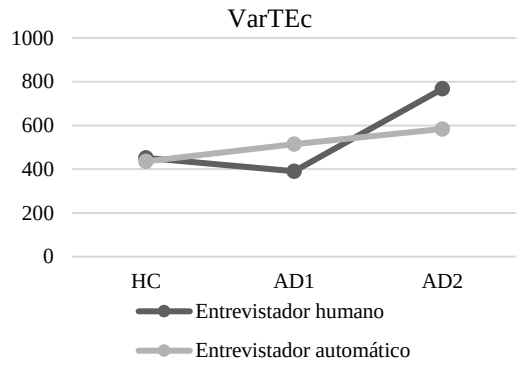


Figura 9-12 Estadísticos descriptivos: comparación de los valores de VarTEc ($\sigma^2_{te_c}$) para entrevistador humano y automático.

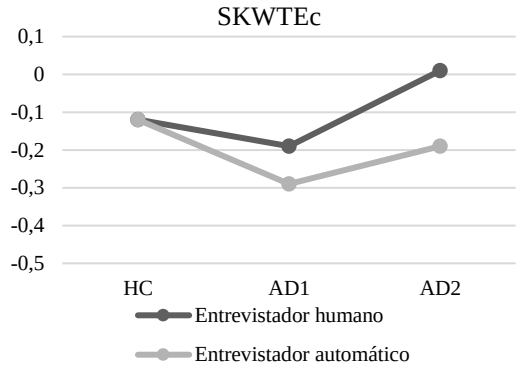


Figura 9-13 Estadísticos descriptivos: comparación de los valores de SKWTEc ($\tilde{\mu}_{te_{c_3}}$) para entrevistador humano y automático.

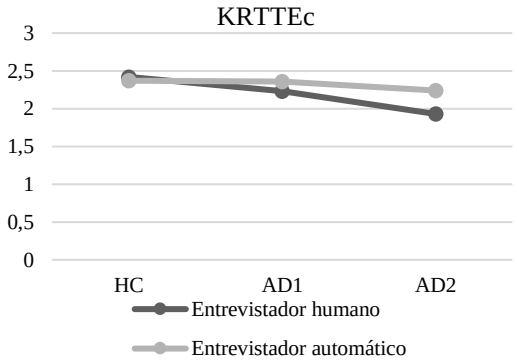


Figura 9-14 Estadísticos descriptivos: comparación de los valores de KRTTEc ($Kurt_{te_c}$) para entrevistador humano y automático.

Por su parte, a partir de los tres análisis no paramétricos realizados: Test de Wilcoxon, Test de Kruskal-Wallis y Test de la Mediana, hemos podido conocer cuáles de estas cinco variables resultarían ser discriminantes AD realizando cuatro comparaciones diferentes entre las poblaciones.

A partir de la Tabla 7-6, las variables que mejores resultados reportan serían, para el entrevistador automático y comparación HC-AD, las variables TEd (te_d) y SKWTEc ($\tilde{\mu}_{te_{c_3}}$) estando, en el caso de la variable TEd (te_d), al límite de los valores que hemos definido como máximos (*Prob|z| ≤ 0,05*). También rechazaríamos la hipótesis nula para la variable SKWTEc ($\tilde{\mu}_{te_{c_3}}$) si comparamos la población HC frente AD1. En cuanto al entrevistador humano, sólo la variable KRTTEc con un *p-value* del 0,05 en los Test de Wilcoxon y Kruskal Wallis rechazaría la hipótesis. Esto significa que la probabilidad de que obtengamos alguna diferencia entre un grupo y otro es del 5% o, dicho de otra manera: sólo hay una probabilidad del 5% de que obtengamos diferencia entre los grupos HC y AD2 si asumimos que todos los grupos son iguales. Por tanto, con los resultados obtenidos podemos decir que se rechaza la hipótesis nula. Las variables MediaTEc ($\overline{te_c}$) y VarTEc ($\sigma^2_{te_c}$) en ninguno de los cuatro escenarios no paramétricos estudiados resultan ser discriminantes. Estos resultados no implican que estas variables no sean discriminantes sino simplemente que no podríamos rechazar la hipótesis nula.

En el caso de la comparación AD1 y AD2 ningún resultado ha quedado por encima del 0,05 para el *p-value*. Como ocurrió en los análisis paramétricos realizados para las variables temporales, probablemente esto se deba a que las poblaciones AD1 y AD2 constituyen conjuntos de muestras no lo suficientemente grandes y, además, muy similares entre sí.

9.4.2 Análisis multivariante

El análisis multivariante realizado ha sido la prueba de la varianza MANOVA. Los resultados recogidos en la Tabla 7-7 muestran cómo, para todos los estadísticos analizados, el set de características utilizado es discriminante AD cuando las muestras analizadas son las del entrevistador automático y, además, cuando lo que se compara son las poblaciones HC y AD.

Sin embargo, cuando entran en juego las variables de agrupación AD1 y AD2, para ambos entrevistadores los valores de los estadísticos están muy por encima del límite máximo establecido para el *p-value*. Éste, en parte, no es un hecho aislado y podría ir en línea con los resultados obtenidos en las pruebas no paramétricas anteriores donde el caso en el que un mayor número de variables cumplió el límite del *p-value* fue en la comparación HC-AD. Como ya se comentó en el apartado anterior, en los tres test univariantes no paramétricos llevados a cabo, ninguna de las cinco variables por separado resultaba ser discriminante cuando lo que comparábamos eran las poblaciones AD1 y AD2. El motivo podría ser el expuesto anteriormente sobre el número de muestras disponibles para cada población (a mayor número de muestras obtendremos resultados más concluyentes) a lo que se sumaría una mayor dificultad a la hora de diferenciar dos voces patológicas de diferentes grados.

Como puede deducirse a partir de los datos obtenidos, las muestras que rechazarían la hipótesis nula en cuanto a discriminación AD son las de habla inducida, aquellas obtenidas con el entrevistador automático (al contrario que ocurrió con el análisis MANOVA realizado sobre las variables temporales).

9.5 Análisis estadístico multivariante: medidas temporales y Temperatura Emocional

El último de los análisis estadísticos realizados ha sido el MANOVA aplicado al conjunto completo de diez características (temporales y de Temperatura Emocional).

Los resultados recogidos en la Tabla 7-8 muestran cómo, para todos los estadísticos analizados, el set de características rechazaría la hipótesis nula para los dos tipos de entrevistadores y para todas las comparaciones entre poblaciones HC-AD, y HC-AD1 y HC-AD2. Sin embargo, cuando las variables de agrupación son las poblaciones AD1 y AD2, para ambos entrevistadores los valores de los estadísticos están muy por encima del límite establecido para el *p-value* (0,05). Éste, en cualquier caso, no es un hecho aislado e iría en línea con los resultados obtenidos en otras pruebas anteriores, como los estudios no paramétricos realizados. En los tres test univariantes no paramétricos llevados a cabo tanto para el conjunto de medidas temporales como para el conjunto de medidas de Temperatura Emocional, ninguna de las variables por separado resultaba ser discriminantes cuando lo que comparábamos eran las poblaciones AD1 y AD2. De nuevo, el motivo podría ser el expuesto anteriormente sobre el número de muestras disponibles para cada población a lo que se sumaría una mayor dificultad a la hora de diferenciar dos voces patológicas de diferentes grados.

En este análisis, tanto las muestras de habla espontánea como aquellas de habla inducida rechazarían la hipótesis nula. Estos resultados indican que mediante una combinación de características temporales y de Temperatura Emocional podrían distinguirse poblaciones y, por tanto, identificarse mediante un estudio más pormenorizado determinadas características que fueran discriminantes AD.

9.6 Análisis de los clasificadores

9.6.1 Análisis discriminante

Medidas temporales

El análisis multivariante por medio de clasificadores en Stata lo hemos basado en el uso de tres modelos diferentes: LDA, clasificador logístico y clasificador kNN. Si hacemos un análisis atendiendo a los tres clasificadores indicados, vemos que existe una clara diferencia entre los resultados obtenidos para el clasificador LDA (valores medios: tasa de éxito = 82,6%, sensibilidad = 84,1%, especificidad = 78,2%) y el clasificador logístico (valores medios: tasa de éxito = 86,1%, sensibilidad = 87,6%, especificidad = 85,9%) frente al clasificador kNN en cualquiera de sus variantes (*n = 1*, *n = 3* o *n = 5*) (valores medios: tasa de éxito = 59,8%, sensibilidad = 62%, especificidad = 56,7%).

A partir de la Figura 9-15, Figura 9-16, Figura 9-17, Figura 9-18, Figura 9-19 y Figura 9-20 obtenidas de la Tabla 8-1, Tabla 8-2, Tabla 8-3 y Tabla 8-4 podemos observar que, tanto el clasificador LDA como el logístico, ofrecen los mejores resultados siendo, a su vez, el análisis logístico el óptimo en cuanto a prestaciones.

Si nos fijamos en las poblaciones estudiadas, hemos realizado dos tipos de clasificaciones: por presencia o ausencia de enfermedad (HC-AD) y diferenciando por grados (HC-AD1-AD2). Los valores de los diferentes clasificadores cuando realizamos ambas agrupaciones son muy similares y reflejan una tendencia más o menos general. Sin embargo puede observarse que el primer caso, en el que se comparan la población sana con la patológica, ofrece resultados ligeramente mejores que la clasificación por grados. Concretamente, los clasificadores LDA y logístico para la clasificación por presencia o ausencia de enfermedad tenemos una tasa de éxito del 85,8% y 92,9% para el entrevistador humano y del 80% y 79,4% para el entrevistador automático. Estos valores son superiores en, aproximadamente, un 1% si los comparamos con sus homólogos en la clasificación por grados de enfermedad. Esta variación podría deberse a la complicación añadida de diferenciar entre dos voces que sean patológicas de diferentes grados y que, por tanto, presentan similitudes temporales. En cualquier caso, sí que podemos decir que la afirmación anterior por la que los clasificadores que mejores resultados arrojan son el clasificador LDA y logístico, se cumple tanto para la clasificación biclase como para la clasificación por grados de enfermedad. En cuanto al análisis de los resultados atendiendo al entrevistador utilizado, en general, como hemos dicho, el entrevistador humano obtiene mejores resultados, si bien es cierto que las diferencias entre ambos entrevistadores se estrechan considerablemente en algunos casos.

Si tomamos como referencia el escenario que mejores resultados reporta, el clasificador logístico sobre las poblaciones HC y AD, tenemos que la tasa de éxito del entrevistador humano es de un 92,9% frente al 79,4% del entrevistador automático. Respecto a la sensibilidad y especificidad tenemos que el entrevistador humano alcanza un 94% y 93%, respectivamente, frente al 85% y 75% obtenidos con el entrevistador automático. En términos diferenciales esto supone una reducción de los resultados del entrevistador automático frente al humano del 14%, 9% y 18%, respectivamente. Aunque estas diferencias son considerables, los valores ofrecidos por el entrevistador automático resultan nada desdeñables, especialmente teniendo en cuenta que, la automatización de los procesos de entrevistas aún supone un campo por explorar.

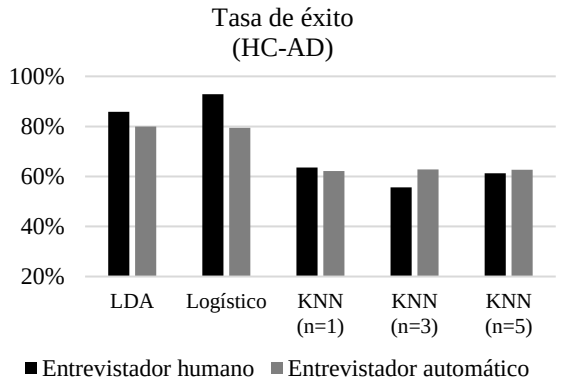


Figura 9-15 Análisis discriminante: clasificación por presencia o ausencia AD. Comparación de los valores de tasa de éxito de los diferentes clasificadores para el entrevistador humano y automático (medidas temporales).

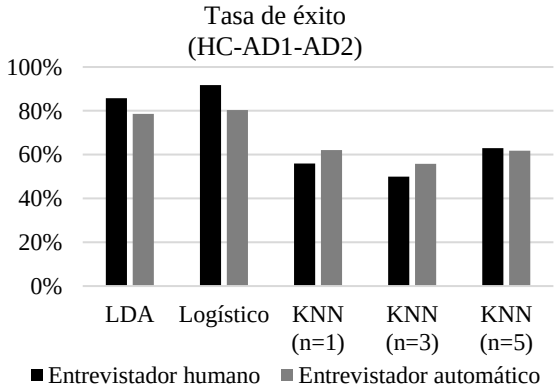


Figura 9-16 Análisis discriminante: clasificación por grados AD. Comparación de los valores de tasa de éxito de los diferentes clasificadores para el entrevistador humano y automático (medidas temporales).

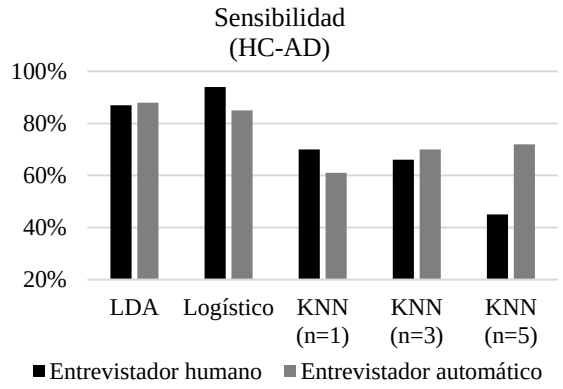


Figura 9-17 Análisis discriminante: clasificación por presencia o ausencia AD. Comparación de los valores de sensibilidad de los diferentes clasificadores para entrevistador el humano y automático (medidas temporales).

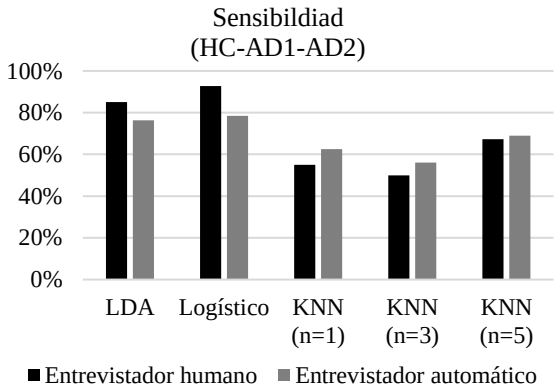


Figura 9-18 Análisis discriminante: clasificación por grados AD. Comparación de los valores de sensibilidad de los diferentes clasificadores para entrevistador el humano y automático (medidas temporales).

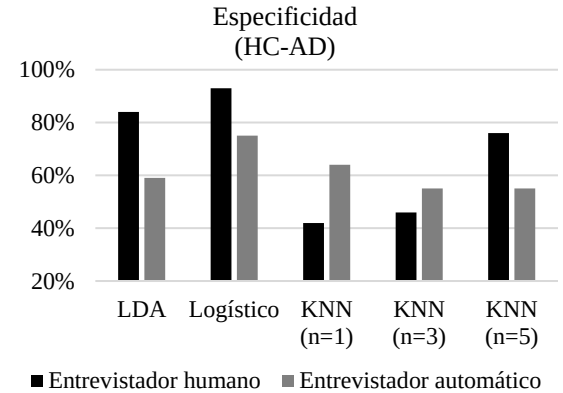


Figura 9-19 Análisis discriminante: clasificación por presencia o ausencia AD. Comparación de los valores de especificidad de los diferentes clasificadores para el entrevistador humano y automático (medidas temporales).

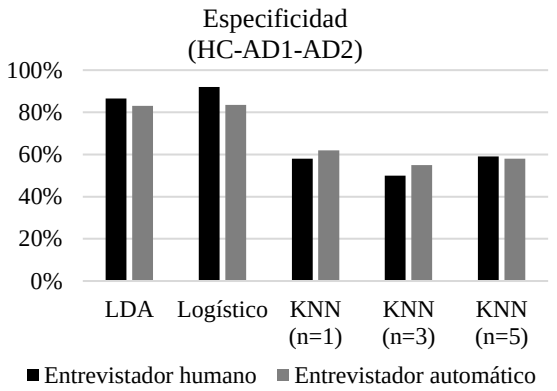


Figura 9-20 Análisis discriminante: clasificación por grados AD. Comparación de los valores de especificidad de los diferentes clasificadores para el entrevistador humano y automático (medidas temporales).

Temperatura Emocional

Como en el caso anterior, este tipo de análisis multivariante mediante clasificadores lo hemos basado en el uso de tres clasificadores diferentes: LDA, clasificador logístico y clasificador kNN (*n = 1*, *n = 3* y *n = 5*).

A simple vista a partir de la Figura 9-21, Figura 9-22, Figura 9-23, Figura 9-24, Figura 9-25 y Figura 9-26 obtenidas de la Tabla 8-5, Tabla 8-6, Tabla 8-7 y Tabla 8-8, si hacemos un análisis atendiendo estos tres clasificadores, vemos que para los valores de tasa de éxito y especificidad los mejores resultados se obtienen para la clasificación biestado (sano-patológico). La tasa de éxito alcanza su valor máximo con el clasificador logístico tanto para el entrevistador automático (tasa de éxito = 59,4%) como para el humano (tasa de éxito = 57,8%). Respecto a los valores de especificidad, su máximo valor se logra con el entrevistador humano (especificidad = 69,6%) frente al automático (especificidad = 62,3%) y, en ambos casos, mediante el clasificador kNN (*n = 3* y *n = 5*, respectivamente).

Volviendo a la comparación entre los dos tipos de clasificaciones (enfermedad/grados) vemos que, al contrario que ocurre con los valores de tasa de éxito y especificidad, los valores para la sensibilidad son mejores en la clasificación por grados. Existe una clara diferencia en favor de este último tipo de clasificación cuyos mejores valores se sitúan por encima del 70% tanto para el entrevistador humano (sensibilidad = 72% para el clasificador logístico y kNN, $n = 3$) y para el automático (sensibilidad = 73,9% para el clasificador kNN con $n = 5$). En cualquier caso, cabe destacar los bajos valores obtenidos sea cual sea el clasificador empleado. Concretamente, en el caso en que la sensibilidad alcanza su máximo con el entrevistador automático, en contraposición, encontramos un valor de especificidad que no es aceptable (especificidad = 26,1%).

A pesar de que, dada la naturaleza de nuestros objetivos, el parámetro más interesante para nosotros sería la sensibilidad, es decir la capacidad de identificar un sujeto patológico y no lo contrario, un valor tan bajo en la especificidad nos indica un exceso inasumible de falsos positivos en la clasificación. Respecto al sentido de la tasa de éxito, que representa la dispersión de los valores de las muestras, hace referencia a la proporción entre el número de predicciones correctas (tanto positivas como negativas) y el total de predicciones. Conceptualmente, el "coste" asociado a cada error de clasificación del algoritmo es lo que debemos considerar a la hora de centrarnos más o menos en un parámetro. Dado que en la cuestión planteada es muy importante detectar sujetos patológicos, la sensibilidad debe ser el principal parámetro a la hora de estimar nuestros clasificadores, si bien es cierto que es necesario un compromiso entre sensibilidad-especificidad (centrada esta última en la identificación de casos negativos clasificados correctamente) para que nuestro sistema esté realmente equilibrado.

De manera general, hemos encontrado que los clasificadores con los mejores resultados son kNN ($n = 3$ y $n = 5$) y clasificador logístico. Para los valores de sensibilidad, ambos entrevistadores obtienen sus mejores resultados para la clasificación por grados, con valores en torno al 70%. En cuanto al análisis de los resultados atendiendo al entrevistador utilizado, en general, no existe una ventaja clara de uno sobre otro para ninguno de los parámetros analizados.

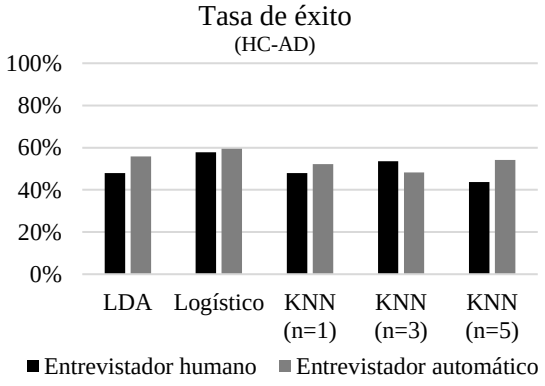


Figura 9-21 Análisis discriminante: clasificación por presencia o ausencia AD. Comparación de los valores de tasa de éxito de los diferentes clasificadores para el entrevistador humano y automático (Temperatura Emocional).

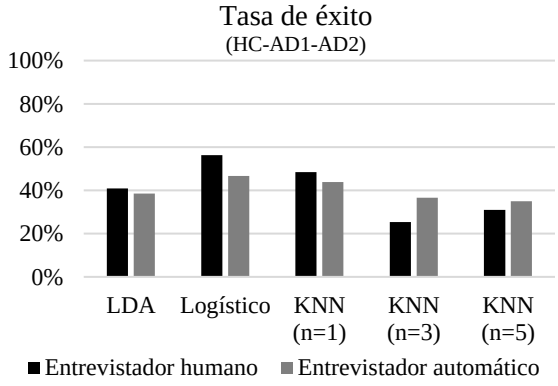


Figura 9-22 Análisis discriminante: clasificación por grados AD. Comparación de los valores de tasa de éxito de los diferentes clasificadores para el entrevistador humano y automático (Temperatura Emocional).

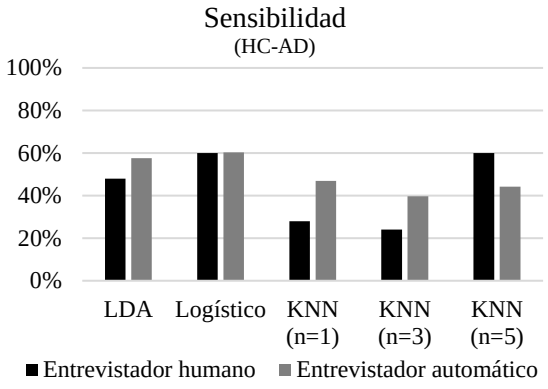


Figura 9-23 Análisis discriminante: clasificación por presencia o ausencia AD. Comparación de los valores de sensibilidad de los diferentes clasificadores para entrevistador el humano y automático (Temperatura Emocional).

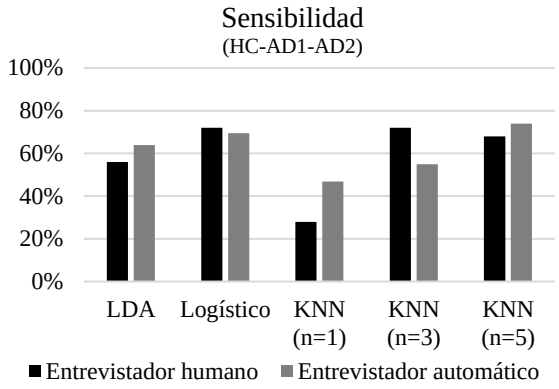


Figura 9-24 Análisis discriminante: clasificación por grados AD. Comparación de los valores de sensibilidad de los diferentes clasificadores para entrevistador el humano y automático (Temperatura Emocional).

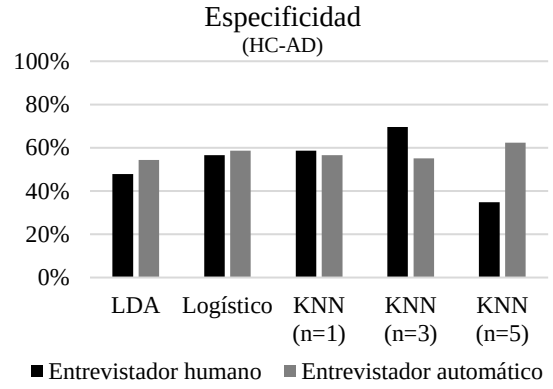


Figura 9-25 Análisis discriminante: clasificación por presencia o ausencia AD. Comparación de los valores de especificidad de los diferentes clasificadores para el entrevistador humano y automático (Temperatura Emocional).

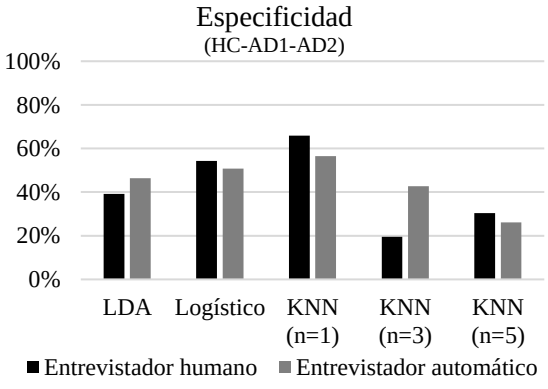


Figura 9-26 Análisis discriminante: clasificación por grados AD. Comparación de los valores de especificidad de los diferentes clasificadores para el entrevistador humano y automático (Temperatura Emocional).

Medidas temporales y de Temperatura Emocional

El último de los análisis con clasificadores que hemos llevado a cabo con el *software* Stata se ha realizado partiendo del set completo de características, tanto temporales como de Temperatura Emocional. El análisis se ha basado, como se hiciera en los casos anteriores, en el uso de cuatro clasificadores diferentes: LDA, clasificador logístico y clasificador kNN.

De manera global puede apreciarse una mejora de los resultados con respecto a los clasificadores aplicados exclusivamente sobre las características de Temperatura Emocional. Para los tres estimadores analizados (tasa de éxito, sensibilidad y especificidad) y prácticamente en todos los escenarios estudiados, los mejores valores se han obtenido para el clasificador logístico que, en todos los casos, ha sido mejor para el entrevistador humano (tasa de éxito = 97,2%, sensibilidad = 96% y especificidad = 97,8% en la clasificación por grados) que para el automático (tasa de éxito = 80,7% y especificidad = 86,2% en la clasificación por enfermedad y sensibilidad = 76,6% en el caso de clasificación por grados). Sin embargo, tampoco se aprecia, de manera general,

una reducción significativa de los valores obtenidos para la clasificación por grados de enfermedad respecto a la clasificación por ausencia o presencia de la misma.

Por su parte, a partir de la Figura 9-27, Figura 9-28, Figura 9-29, Figura 9-30, Figura 9-31 y Figura 9-32 obtenidas de la Tabla 8-9, Tabla 8-10, Tabla 8-11 y Tabla 8-12, podemos observar que, prácticamente en todos los casos, los valores del entrevistador humano superan a los del entrevistador automático, siendo las diferencias entre uno y otro, a menudo, significativas (valores del entrevistador automático rondando una baja de aproximadamente el 10% del valor obtenido para el entrevistador humano).

Para tener una idea general de todos los resultados obtenidos del análisis discriminante, hemos hecho un repaso de todos los escenarios estudiados por tipo de clasificación, medidas utilizadas, entrevistador y clasificador. Hemos seleccionado los diez mejores resultados de sensibilidad (parámetro que para nuestro caso concreto resulta de especial interés). Los mejores resultados a nivel global se muestran en la Tabla 9-1.

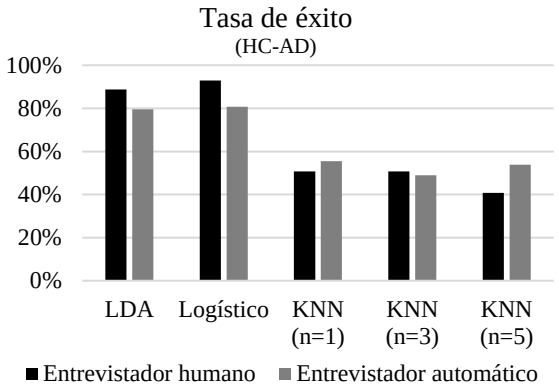


Figura 9-27 Análisis discriminante: clasificación por presencia o ausencia AD. Comparación de valores de tasa de éxito de los diferentes clasificadores para el entrevistador humano y automático (medidas temporales y de Temperatura Emocional).

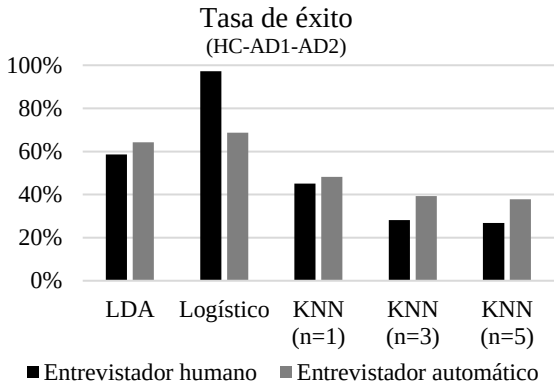


Figura 9-28 Análisis discriminante: clasificación por grados AD. Comparación de valores de tasa de éxito de los diferentes clasificadores para el entrevistador humano y automático (medidas temporales y de Temperatura Emocional).

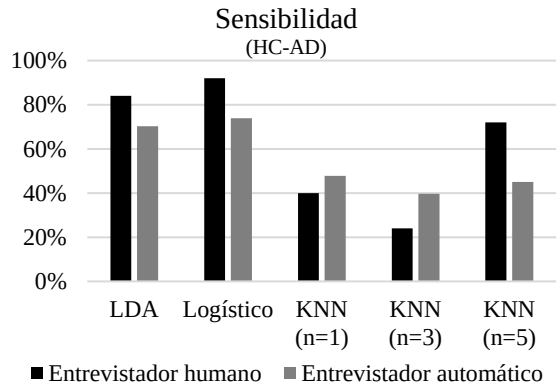


Figura 9-29 Análisis discriminante: clasificación por presencia o ausencia AD. Comparación de los valores de sensibilidad de los diferentes clasificadores para entrevistador el humano y automático (medidas temporales y de Temperatura Emocional).

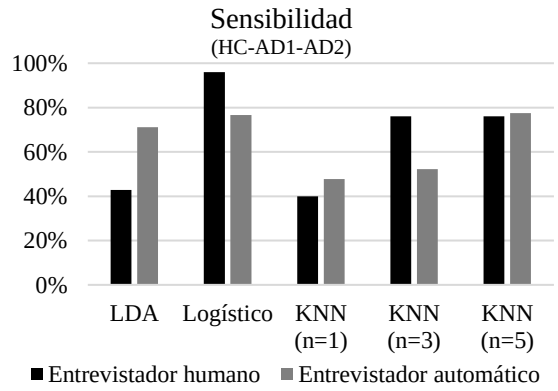


Figura 9-30 Análisis discriminante: clasificación por grados AD. Comparación de los valores de sensibilidad de los diferentes clasificadores para entrevistador el humano y automático (medidas temporales y de Temperatura Emocional).

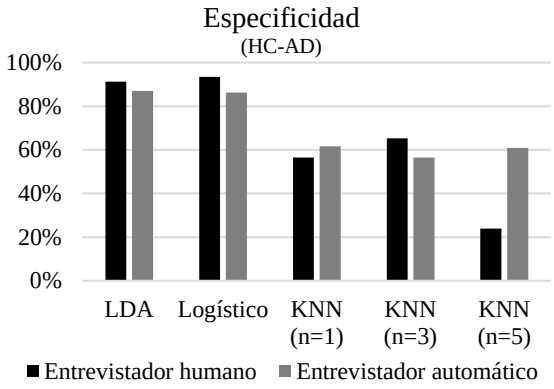


Figura 9-31 Análisis discriminante: clasificación por presencia o ausencia AD. Comparación de valores de especificidad de diferentes clasificadores para entrevistador humano y automático (medidas temporales y de Temperatura Emocional).

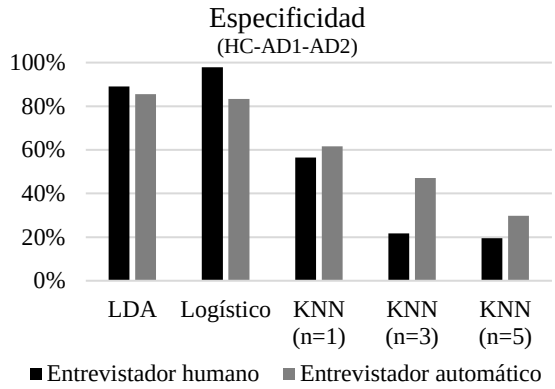


Figura 9-32 Análisis discriminante: clasificación por grados AD. Comparación de valores de especificidad de los diferentes clasificadores para el entrevistador humano y automático (medidas temporales y de Temperatura Emocional).

Tabla 9-1 Análisis discriminante: selección de los diez mejores clasificadores a partir de los valores de sensibilidad.

Set medidas	**Clasificación**	**Entrevistador**	**Clasificador**	**ACC [%]**	**Sensibilidad [%]**	**Especificidad [%]**
Temporales	Enfermedad	Automático	LDA	80,00%	88,00%	59,00%
Temporales	Enfermedad	Automático	Logístico	79,40%	85,00%	75,00%
Temporales	Enfermedad	Humano	LDA	85,80%	87,00%	84,00%
Temporales	Enfermedad	Humano	Logístico	92,90%	94,00%	93,00%
Temporales	Grados	Automático	Logístico	80,42%	78,50%	83,50%
Temporales	Grados	Humano	LDA	85,79%	85,14%	86,60%
Temporales	Grados	Humano	Logístico	91,80%	92,80%	92,00%
Temporales + TE	Enfermedad	Humano	LDA	88,73%	84,00%	91,30%
Temporales + TE	Enfermedad	Humano	Logístico	92,96%	92,00%	93,48%
Temporales + TE	Grados	Humano	Logístico	97,18%	96,00%	97,83%

De la Tabla 9-1 podemos extraer varias ideas. La primera es que, claramente, los sets que mejores resultados reportan son los que comprenden las medidas temporales y el conjunto completo: temporales y de Temperatura Emocional. Concretamente, utilizando la combinación de ambos tipos de características se obtendría el mejor resultado (clasificación por grados, entrevistador humano y clasificador logístico: tasa de éxito = 97,2%, sensibilidad = 96% y especificidad = 97,8%). El uso exclusivo del set de características de Temperatura Emocional, al menos bajo las condiciones en las que se ha realizado este estudio, no conseguiría estar entre los mejores resultados de clasificación.

Es importante mencionar que para la escala de colores empleada en la Tabla 9-1 se han tomado como referencia exclusivamente los que hemos considerado los diez mejores clasificadores. Teniendo en cuenta este hecho, resulta especialmente interesante que, de las 60 configuraciones analizadas en este apartado (diferentes conjuntos de medidas, clasificaciones y modelos), tres clasificadores automáticos se encuentren entre los diez mejores valores de sensibilidad.

Respecto al tipo de clasificación empleada, la mayor parte de estos clasificadores seleccionados como mejores se basa en una clasificación por presencia/ausencia de enfermedad. Al mismo tiempo también es cierto que las

clasificaciones por grados se encuentran entre las mejores. En este sentido, con los datos que disponemos, no podemos extraer una conclusión clara al respecto.

A continuación se presenta la Tabla 9-2 con una comparación de los mejores resultados obtenidos para el entrevistador humano y automático.

Tabla 9-2 Análisis discriminante: comparativa entre los mejores resultados obtenidos para el entrevistador humano y automático.

Entrevistador	Set medidas	Clasificación	Clasificador	ACC [%]	Sensibilidad [%]	Especificidad [%]
Automático	Temporales	Enfermedad	LDA	80,00%	88,00%	59,00%
Automático	Temporales	Enfermedad	Logístico	79,40%	85,00%	75,00%
Humano	Temporales + TE	Grados	Logístico	97,18%	96,00%	97,83%

Uno de los que presentamos como principales objetos de esta tesis se centraba en arrojar luz sobre las posibilidades o no que el entrevistador automático pudiera tener. A partir de la Tabla 9-2, efectivamente se aprecia que los peores resultados se obtienen para el entrevistador automático. En el caso de la Tabla 9-2, encontramos que el mejor clasificador para el caso del entrevistador automático se daría, en términos de sensibilidad, para la configuración basada en medidas: temporales, clasificación: enfermedad y clasificador: LDA. Sin embargo, podemos comprobar que el valor para la especificidad cae a niveles poco recomendables (especificidad = 59%). Dado que hemos considerado que es necesario alcanzar un compromiso entre sensibilidad y especificidad, se ha incluido en esta tabla otro de los mejores clasificadores para el entrevistador automático: la configuración basada en medidas temporales, clasificación: enfermedad y clasificador: logístico. En este caso, si bien la sensibilidad cae en tres puntos porcentuales respecto a la configuración anterior, el valor de la especificidad aumenta notablemente (especificidad = 75%), compensando este hecho mediante una mayor detección de sujetos sanos.

9.6.2 Modelos de clasificación

El estudio realizado con modelos de clasificación en Matlab se ha llevado a cabo en diferentes fases. En primer lugar se ha realizado la selección de características considerando la afección que pueda tener el tipo de entrevistador con el que se han obtenido las muestras, así como la influencia del número de poblaciones utilizadas. A partir de esa comprobación, se ha establecido como criterio de selección la relevancia de las características tomando como referencia el entrevistador automático y clasificación biestado.

Los resultados obtenidos han reportado que las características con más peso serían: VarHabla ($\sigma_{t_s}^2$), KRTHabla ($Kurt_{t_s}$), INDHabla (Ind_{t_s}), TEd (te_d), MediaTEc ($\overline{te_c}$) y VarTEc ($\sigma_{te_c}^2$). Siendo INDHabla (Ind_{t_s}) y MediaTEc ($\overline{te_c}$), las que ofrecen los mejores resultados para cada conjunto temporal o emocional, respectivamente. La primera de ellas avalada, también, por el estudio univariante realizado en el Capítulo 7.

A partir de las características identificadas como más relevantes (ver apartado 8.2.1), el siguiente paso ha consistido en conocer cuál es el conjunto óptimo de cara a lograr los mejores resultados de clasificación. A partir de las combinaciones posibles de

estas seis mejores características y de un clasificador SVM como referencia (ver apartado 0), se ha procedido a calcular los valores de exactitud, sensibilidad y especificidad para cada posible combinación, tanto para las diez medidas temporales y de Temperatura Emocional, como para diferentes combinaciones de ellas. Como resultado se llega a tres conjuntos óptimos posibles:

- CO1: VarHabla $\left(\sigma_{t_S}^2\right)$, KRTHabla $(Kurt_{t_s})$ e INDHabla (Ind_{t_s}).
- CO2: VarHabla $\left(\sigma_{t_S}^2\right)$, KRTHabla $(Kurt_{t_s})$ e INDHabla $\left(Ind_{t_s}\right)$ y MediaTEc $(\overline{te_c})$.
- CO3: VarHabla $\left(\sigma_{t_S}^2\right)$, KRTHabla $(Kurt_{t_s})$ e INDHabla $\left(Ind_{t_s}\right)$, TEd (te_d), MediaTEc $(\overline{te_c})$ y VarTEc $(\sigma_{te_c}^2)$.

La última parte se ha centrado en el estudio de los resultados también en términos de exactitud, sensibilidad y especificidad obtenidos de aplicar las mejores combinaciones de características a seis modelos de clasificación optimizados. A saber: *tree*, *discriminant*, kNN, *Naive Bayes*, SVM y *ensemble* (ver apartado 0). Para estas simulaciones se ha analizado el caso de que la respuesta esté basada tanto en la variable Enfermedad (HC, AD) como en la variable Grados (HC, AD1, AD2).

De todos los resultados obtenidos, en la Tabla 9-3 se muestran los que mejores resultados han obtenido en términos de tasa de éxito, sensibilidad y especificidad.

Tabla 9-3 Clasificadores: selección de los mejores clasificadores en Matlab.

Clasificación	Entrevistador	Clasificador	Conj. óptimo	ACC [%]	S [%]	E [%]
Enfermedad	**Automático**	**Ensemble**	CO1	82,00	77,00	87,00
Grados	**Automático**	**Ensemble**	CO1	75,00	70,00	93,00
Enfermedad	**Humano**	**SVM**	CO1	92,00	84,00	96,00
Grados	**Humano**	**SVM**	CO1	85,00	84,00	91,00

A continuación, en la Figura 9-33 pueden analizarse gráficamente estos resultados. Puede comprobarse que, con diferencia, el entrevistador humano obtiene mejor rendimiento, tal y como reportaron los resultados de los clasificadores implementados en Stata.

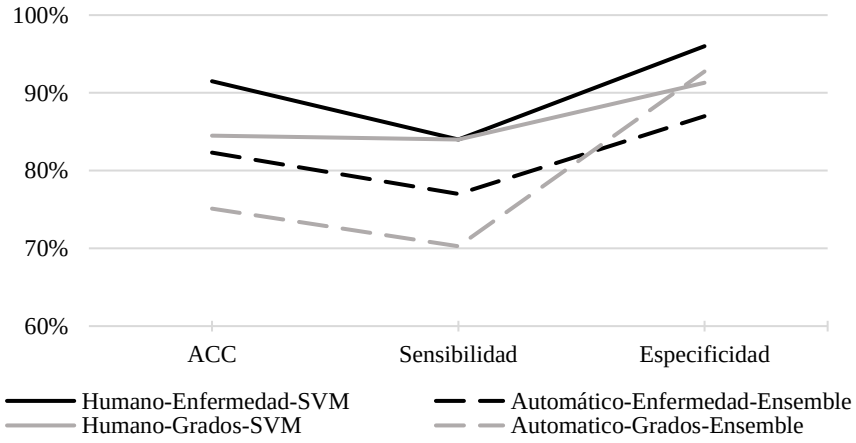


Figura 9-33 *Classification-Learner App* Matlab: selección de los cuatro modelos para entrevistador humano y automático con mayor rendimiento (tasa de éxito, sensibilidad y especificidad).

A partir de la Figura 9-33 podemos ver que la clasificación por grados empeora también los resultados para ambos entrevistadores, lo que nos lleva a plantear que, como se dijo anteriormente, con la base de datos actual, donde el número de muestras para las clases leves y moderadas es menor que el de muestras HC, el rendimiento de los clasificadores se reduce considerablemente.

Por otro lado, podemos apreciar que el entrevistador humano consigue sus mejores resultados utilizando un clasificador SVM mientras que el automático optimiza sus marcas con modelos tipo *ensemble* (modelos AdaBoost y *Bagged Tree*). Otro hecho reseñable es que el conjunto de características óptimo para cada escenario se obtiene realizando una importante selección de características: reduciendo de diez características en total, a tan sólo tres variables temporales.

Sobre las variables que se han escogido como mejores a partir del proceso de selección de características, cabe decir que irían aproximadamente en línea con los resultados obtenidos del análisis univariante realizado en Stata.

En este primer análisis, fundamentalmente a partir de los estadísticos descriptivos, pudo extraerse que las mejores variables temporales en la discriminación de la enfermedad eran MediaHabla ($\bar{t_S}$), VarHabla $\left(\sigma_{t_S}^2\right)$ e INDHabla (Ind_{t_s}). Estas variables aportaban información de la duración de los tiempos de habla de los sujetos, siendo mayores en los sujetos de control y los pacientes AD leves que en los pacientes AD moderados, quienes tienen afectadas las capacidades comunicativas en mayor medida. Por su parte, no se obtuvieron, a partir de ese análisis, resultados concluyentes respecto a las variables SKWHabla ($\tilde{\mu}_{t_{S_3}}$) y KRTHabla ($Kurt_{t_s}$).

Del proceso de selección de características realizado en Matlab, hemos podido coincidir con el primer análisis univariante en que dos de ellas, VarHabla $\left(\sigma_{t_S}^2\right)$ e INDHabla (Ind_{t_s}), sí que son discriminantes mientras que MediaHabla ($\bar{t_S}$), perdería relevancia en pro de la variable KRTHabla ($Kurt_{t_s}$). La interpretación que podemos hacer de este hecho es que, analizadas por separado, sin influencia de otras variables, MediaHabla ($\bar{t_S}$) puede ser un buen indicador, sin embargo, considerada junto con otras variables, su relevancia pierde peso. Al contrario que habría pasado con la variable KRTHabla ($Kurt_{t_s}$), que en el proceso de seleccion de características sería una de las más significativas.

Por último, hemos querido comparar los resultados obtenidos del análisis discriminante realizado en Stata (importante reseñar que en este caso no se ha hecho selección de características) con los de los modelos entrenados en Matlab (que, además de incluir el conjunto total de características, ha sido sometido al ya mencionado proceso de selección de características).

El mejor clasificador obtenido del análisis discriminante ha sido, para el entrevistador humano, el modelo logístico con clasificación por grados de enfermedad y uso de medidas temporales combinadas con medidas de Temperatura Emocional (tasa de éxito = 97,2%, sensibilidad = 97,8% y especificidad = 96%). En el caso de modelos en Matlab, para el entrevistador humano, el mejor clasificador ha sido SVM, con el conjunto óptimo CO1 y clasificación por presencia o ausencia de enfermedad (tasa de éxito = 91,5%, sensibilidad = 84% y especificidad = 96%).

Por otra parte, centrándonos en el entrevistador automático, en la Tabla 9-4 podemos encontrar los resultados de los mejores modelos (Stata-M1, Stata-M2 y Matlab-M3).

Tabla 9-4 Entrevistador automático: selección de los mejores clasificadores mediante análisis discriminante Stata y *Classification Learner* en Matlab.

Denominación	Medidas utilizadas	Clasificación	Entrevistador	Clasificador	ACC [%]	S [%]	E [%]
Stata – M1	Temporales	Enfermedad	Automático	LDA	80,00	88,00	59,00
Stata – M2	Temporales	Enfermedad	Automático	Logístico	79,00	85,00	75,00
Matlab – M3	CO1	Enfermedad	Automático	Ensemble	82,00	77,00	87,00
CO1: conjunto óptimo de características 1 (ver apartado 8.2.1).							

Por un lado tenemos que el clasificador LDA (Stata-M1) ofrece los mejores valores de sensibilidad (parámetro que, como habíamos indicado, es el más interesante para el caso planteado). Sin embargo, este clasificador a pesar de tener los mejores resultados en términos de tasa de éxito y sensibilidad (tasa de éxito = 80%, sensibilidad = 88%), también tiene valores de especificidad muy bajos (especificidad = 59%). Esto indica que se clasificarán como positivas muchas muestras que realmente corresponden a sujetos HC.

Por otro lado, el siguiente mejor clasificador (Stata-M2), modelo logístico y clasificación por enfermedad, ofrece una sensibilidad y tasa de éxito ligeramente menor que el anterior, si bien la especificidad aumenta considerablemente respecto al modelo Stata-M1 (tasa de éxito = 79,4%, sensibilidad = 85%, especificidad = 75%).

Como ya se ha comentado anteriormente, el estimador más relevante para nosotros es, sin duda, la sensibilidad. Sin embargo, en todos los casos es necesario alcanzar un compromiso entre sensibilidad y especificidad para garantizar un buen funcionamiento de los modelos, especialmente si lo que buscamos a largo plazo es desarrollar herramientas de ayuda al diagnóstico. Concretamente, una especificidad del 59% resulta excesivamente baja a pesar de tener una sensibilidad del 88%. En este caso hablamos de un alto coste asociado a la baja especificidad y derivado de incluir como patológicas muestras que pertenecerían a sujetos sanos.

Cuando pasamos a comparar estos clasificadores con los obtenidos por Matlab y el proceso de selección de características descrito, encontramos que la sensibilidad se reduce considerablemente (del 88% obtenido con Stata-M1 hasta el 77% obtenido con Matlab-M3). El valor de tasa de éxito, sin embargo, es el más alto de los tres escenarios mostrados en la Tabla 9-4. Esto se debe, fundamentalmente, a que los valores de especificidad son mucho mayores para Matlab-M3 que para cualquiera de los otros dos clasificadores implementados con Stata. Ese aumento sustancial en la especificidad (del 59 y 75% de los dos modelos Stata, al 87% obtenido por Matlab-M3), repercute directamente sobre la tasa de éxito, alcanzando un mayor equilibrio entre el número de muestras sanas correctamente clasificadas y el número de muestras patológicas clasificadas como tal. En la Figura 9-34 se representa gráficamente los valores obtenidos para cada escenario.

A partir de la Figura 9-34 podemos observar que los tres modelos mantienen valores similares de tasa de éxito a costa de variar ampliamente, cada uno de ellos, sus

valores de sensibilidad y especificidad. Cabe destacar que, de nuevo, las características temporales resultan ser las más relevantes.

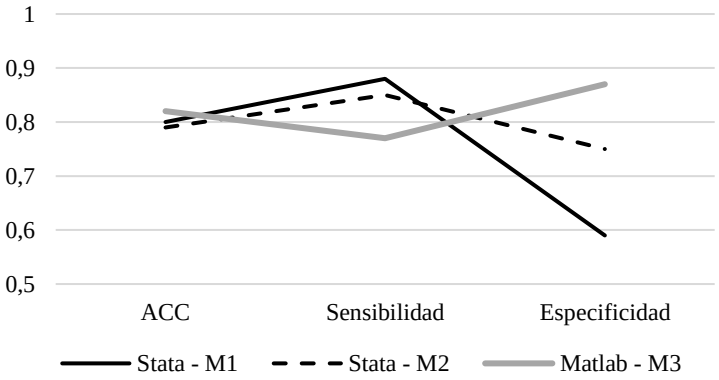


Figura 9-34 Comparativa: análisis discriminante Stata y *Classification-Learner* Matlab: selección de modelos con mayor rendimiento (tasa de éxito, sensibilidad y especificidad). Entrevistador automático.

Para terminar, hemos querido hacer una última comparación en la que se vuelve a uno de los ejes centrales de esta tesis. En este trabajo se han realizado múltiples análisis y estudios desde diferentes perspectivas, pero siempre sin perder de vista lo que ha sido uno de los ejes vertebradores de la tesis: profundizar más en la capacidad discriminante del entrevistador automático y su potencial.

Por este motivo, volviendo al planteamiento inicial, hemos querido reflejar los resultados obtenidos en términos del mejor clasificador humano frente al mejor clasificador automático. Estos resultados se muestran en las Figura 9-35 y Figura 9-36 atendiendo a la herramienta con la que se han obtenido.

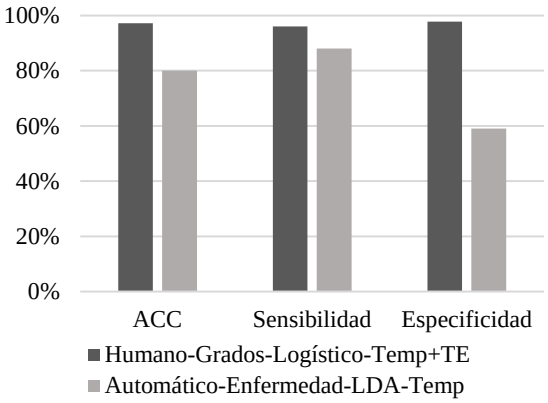


Figura 9-35 Comparativa entre entrevistadores. Análisis discriminante Stata: selección de modelos con mayor rendimiento (tasa de éxito [%], sensibilidad [%] y especificidad [%]).

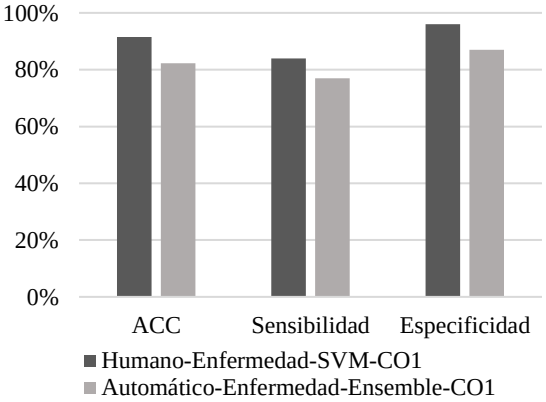


Figura 9-36 Comparativa entre entrevistadores. *Classification Learner* Matlab: selección de modelos con mayor rendimiento (tasa de éxito [%], sensibilidad [%] y especificidad [%]).

Sea cual sea la herramienta de clasificación utilizada, el entrevistador humano muestra mejores resultados. En este punto cabe destacar que los mejores resultados se logran mediante el análisis discriminante en Stata y haciendo uso de las características temporales y de Temperatura Emocional sin aplicar selección de características. Con estas condiciones se alcanza el mejor valor de sensibilidad para el entrevistador automático (sensibilidad = 88%). No obstante, esta misma herramienta muestra unos

estimadores (tasa de éxito, sensibilidad, especificidad) más desbalanceados que los obtenidos mediante la aplicación *Classification Learner* de Matlab.

Por último, fijándonos en la Figura 9-36, podemos observar un hecho bastante significativo y es que las prestaciones (tasa de éxito, sensibilidad, especificidad) siguen la misma tendencia sea cual sea el entrevistador utilizado y que, además, entre uno y otro existe respectivamente una diferencia de 9, 7 y 9 puntos porcentuales. Este hecho pondría de manifiesto comportamientos similares en cuanto a la capacidad discriminante que un entrevistador automático puede tener con respecto a uno humano.

Capítulo 10 Conclusiones y líneas futuras

Desde hace décadas, el número de personas que sufre la enfermedad de Alzheimer en todo el mundo ha aumentado rápidamente. Las cifras son alarmantes y se espera que se tripliquen para 2050. Actualmente, no hay cura y cuando los primeros síntomas son evidentes, el daño causado ya es irreparable y crónico. Debido a los métodos de diagnóstico actuales y la atención y tratamientos a largo plazo, el aumento de la esperanza de vida, así como la inversión de la pirámide generacional, será necesario disponer de sistemas de salud más eficientes y racionales en todo el mundo para hacer frente a esta enfermedad de la que, todavía, no sabemos exactamente cómo, cuándo o por qué se desarrolla.

Es importante aclarar que, aunque todavía hay mucho que aprender, existen multitud de investigaciones en el campo y estrategias de prevención que se están desarrollando actualmente. Hoy, seguramente sabemos más que nunca sobre la AD y esto mejorará con el progreso de la tecnología. Está claro que la detección temprana ayudaría a comprender mejor la enfermedad. Debido al tiempo necesario para diagnosticar la AD, actualmente, cuando los especialistas tratan a los pacientes, es tarde y la efectividad del tratamiento disminuye significativamente. Conocer con mayor exactitud cuándo comienza la enfermedad, incluso en sus fases preclínicas, podría ser fundamental para la efectividad de los tratamientos actuales y, para estudios longitudinales posteriores, especialmente interesante. En base a esto, encontrar biomarcadores accesibles se presenta como una tarea clave para la detección precoz y control evolutivo de la AD.

Recientemente, numerosos estudios han demostrado que el análisis de voz o habla es un poderoso indicador del estado cognitivo de los pacientes con AD. Esto permite detectar los primeros síntomas años antes de que se establezca el diagnóstico clínico probable. Este hecho es bien conocido por todos los expertos en el campo que han administrado manualmente diferentes pruebas neuropsicológicas para analizar la información a partir del habla. En este sentido, la punta de lanza para el procesado de voz aplicado a la detección precoz de AD sería detectar qué pacientes con MCI

evolucionan a AD temprana. Con ese fin, además de tener acceso a las bases de datos utilizadas en el campo, sería especialmente interesante recopilar información longitudinal sobre los pacientes, a lo largo del tiempo y en las diferentes etapas de la enfermedad.

Haciendo un recorrido por las bases de datos localizadas, una de las conclusiones que podemos extraer es que la mayoría de ellas son de carácter transversal. Una de las posibles razones por las que hay menos estudios longitudinales puede deberse a que es más difícil trabajar con los sujetos y garantizar su aceptación cuando el proceso es largo y, aún más, para grupos extendidos. Los recursos y el tiempo requerido son mayores y, hasta hoy, existe una importante falta de criterio común en el tratamiento de estas muestras.

En cuanto a la cantidad de datos y su disponibilidad, debido a que recopilar esta información no es una tarea fácil, las bases de datos no son extensas en la mayoría de los casos y, si son grandes, generalmente no se comparten. Desarrollar y trabajar coordinadamente en bases de datos públicas supondría un impulso para el conocimiento y ayudaría a lograr criterios de recogida y procesamiento de muestras comunes. Para ese propósito, asimismo, se necesitaría evidentemente un presupuesto adecuado, así como el personal cualificado necesario e involucrar a los sujetos a participar en estos estudios.

El progreso tecnológico en las últimas décadas muestra un punto de inflexión en el análisis lingüístico. Los nuevos sistemas automáticos ofrecen resultados más rápidos y con menos esfuerzo que sus homólogos manuales. La automatización del proceso de grabación como tal se presenta como un aspecto muy importante a tener en cuenta.

Hasta la fecha, pocos trabajos se han centrado en aplicar estrategias automáticas, al menos, aplicadas a la AD. Esto incluye, por ejemplo, la automatización de pruebas neuropsicológicas o entrevistas a pacientes que actualmente todavía se administran de manera manual. En este caso, aunque existe evidencia de que al menos una de las bases de datos que hemos localizado realiza entrevistas guiadas con avatares informáticos, poco se sabe sobre los beneficios de automatizar estos procesos. Por tanto, parece interesante conocer en qué medida, para un mismo sujeto, los parámetros obtenidos podrían variar en función del tipo de entrevistador utilizado. El poder extender este tipo de estudios a las nuevas técnicas propuestas en el campo del reconocimiento de voz en las que se plantean ideas basadas en el estudio de muestras obtenidas en entornos ruidosos reales como tertulias, calles, cafés y restaurantes está suscitando un gran interés.

En esa tesis, concretamente se han definido dos conceptos para estudiar los beneficios específicos proporcionados por las técnicas de recogida de muestras automáticas en comparación con sus homólogos manuales: habla inducida (obtenida de un entrevistador automático) y habla espontánea (obtenida de un entrevistador humano). A partir de ellos se ha creado una base de datos denominada Cross-Sectional Alzheimer Prognosis R2019 (CSAP-R19), en la que se han tomado muestras de ambos tipos.

Más allá de automatizar el análisis de algunas pruebas específicas, las investigaciones actuales también han logrado avances importantes en el análisis acústico del habla, tanto para mejorar la accesibilidad como para reducir la carga computacional del procesamiento de datos. De manera amplia, estas investigaciones también han

abierto nuevas e importantes áreas en torno al concepto Health 4.0, donde la tendencia es mejorar los aspectos de seguridad de las personas con demencia mediante dispositivos de detección, así como el desarrollo de herramientas basadas en el tratamiento y problemas terapéuticos para los pacientes y sus cuidadores.

Desarrollar aplicaciones web basadas en biomarcadores lingüísticos para globalizar el control evolutivo y farmacológico a través de la señal de voz supondría con vistas a futuro una solución fácil, rápida, no invasiva y escalable. Este tipo de aplicaciones ofrecería valores objetivos y podría complementar el trabajo de diagnóstico de los especialistas. Ninguna de estas técnicas presentadas requiere, en ningún caso, una amplia infraestructura o disponibilidad de equipo médico: sólo se requieren pruebas verbales y entrevistas con el paciente.

Por su parte, hacer que el proceso *screening* y control evolutivo sea más fácil para los pacientes mediante técnicas transparentes, incluso desde el hogar, ayudará en cualquier caso a reducir el estrés causado por los métodos actuales de diagnóstico. En este punto cabe destacar que en el proceso de grabación con el que se ha recogido nuestra base de datos CSAP-R19 contó con una alta aceptación por parte de los pacientes. Esto se presenta como una ventaja importante de estos métodos automáticos, ya que, de cara al futuro, todas estas condiciones mencionadas seguramente ayudarían a confeccionar bases de datos más extensas. Lo cual sería muy útil, no sólo para el diagnóstico médico, sino también para el control evolutivo, el control farmacológico o la detección precoz de la AD, entre otros.

En esta tesis, los estudios analizados se han clasificado según el tipo de características extraídas para el posterior análisis de voz: convencionales o no convencionales. Las investigaciones son particularmente variadas y aún está pendiente profundizar en el análisis de poblaciones y tareas realizadas para llevar a cado las grabaciones de audio de los sujetos.

Aunque la mayoría de las investigaciones en el campo se han centrado en el análisis de las características convencionales de la señal de voz, las no convencionales son cada vez más significativas. Debido a los aspectos lineales y no lineales de la señal de voz, la aplicación de ambos tipos de características combinadas ofrece resultados más completos. Entre ellos, algunas investigaciones han señalado que cambios cognitivos sutiles en el habla en estados tempranos y preclínicos podrían detectarse mediante fractales. Así, las técnicas más convencionales como ASSA y ERA ya han incluido fractales en su conjunto de características acústicas y estacionarias.

Por su parte, las alteraciones de la respuesta emocional que sufre el paciente son susceptibles de análisis y cuantificaciones adicionales a partir de la señal de voz. Esto podría mejorar significativamente los resultados del proceso de detección y podría proporcionar un criterio universal para la clasificación, independientemente del idioma hablado.

En el proceso de clasificación, las evaluaciones experimentales y estadísticas revelan que el uso de algoritmos de *Machine Learning* para clasificar biomarcadores lingüísticos a través de las declaraciones verbales de personas mayores podría facilitar el diagnóstico clínico de la AD. Aunque los estudios realizados en el campo del Aprendizaje Máquina han producido resultados alentadores, existe también la necesidad

de entrenar estos modelos en conjuntos de datos más grandes. Por su parte, las técnicas actuales de Aprendizaje Profundo, por medio del uso de métodos basados en redes neuronales como pueden ser las CNN, RNN o GCNN, entre otras, ofrecen soluciones interesantes a partir de sistemas de clasificación más complejos para el tratamiento de señales de voz y habla, si bien esa línea no ha sido explorada en esta tesis.

Como ya se mencionó, el estado de la técnica es muy heterogéneo, entre otros, con respecto a las tareas utilizadas en las grabaciones, ya que cada estudio sigue sus propias pautas y criterios. Este es un tema clave porque las características obtenidas de las grabaciones, que miden el deterioro y su significado, varían ampliamente según la tarea que se realice. Es necesario aclarar que este trabajo no ha analizado este punto y profundizar en ello podría ser otra investigación en sí misma. Aunque como primera aproximación hemos realizado una clasificación de los estudios relevantes en cuanto al proceso de extracción de características, una nueva clasificación en detalle con respecto a las tareas realizadas o, por ejemplo, con respecto a los diferentes subtipos de AD, representa claramente una tarea muy necesaria. A este respecto, no sólo deberían clasificarse diferentes tipos de MCI, sino también diferentes patologías y sus correspondientes patrones de voz.

A partir de la revisión crítica del estado del arte realizada y teniendo en cuenta los aspectos mencionados, en esta tesis nos hemos basado en las posibilidades que podría tener una herramienta de apoyo capaz de automatizar el proceso de entrevistas o, en otras palabras: conocer si realmente las muestras obtenidas automáticamente ofrecen una capacidad discriminativa suficientemente buena para distinguir sujetos sanos de patológicos. De ser así, las ventajas de ofrecer entrevistas automatizadas con la misma calidad que las tradicionales, son múltiples. Además de las ya mencionadas, el anonimato que brinda al sujeto que la realiza o la objetividad aportada por un entrevistador que no dependa de las habilidades personales de quien guía la entrevista son ventajas evidentes.

Con esta motivación hemos llevado a cabo un estudio pormenorizado centrado principalmente en el proceso de extracción de características temporales y de Temperatura Emocional y la posterior parametrización de muestras obtenidas manual y automáticamente. Asimismo, hemos realizado diferentes estudios estadísticos univariantes y multivariantes que nos han permitido conocer más sobre la capacidad discriminativa que pueden tener ambos tipos de muestras, habla espontánea o inducida. Se han entrenado también varios modelos de clasificación mediante diferentes herramientas para conocer más en detalle la afección de cada una de las características utilizadas por separado, del conjunto completo o de la combinación de las mejores.

Los resultados esperables pero, ahora sí, justificados, muestran la ventaja del entrevistador humano sobre el automático. Sin embargo, los resultados obtenidos para el este último son prometedores si atendemos a que los valores en términos de tasa de éxito, sensibilidad y especificidad siguen tendencias similares a los del entrevistador humano. En algunos casos alcanzando hasta un 85% de sensibilidad (estimador clave para la tarea que nos hemos propuesto de identificación de sujetos AD) y una especificidad del 75%.

Respecto a las características más relevantes, tenemos que las características temporales aportan una gran cantidad de información si bien, la combinación de características temporales con características de Temperatura Emocional es la que mejores resultados ha obtenido en términos absolutos, alcanzando el entrevistador humano una sensibilidad = 96% y especificidad = 98%.

Si bien es necesario seguir mejorando las marcas del entrevistador automático, con los resultados obtenidos parece claro el potencial que podría llegar a tener. En línea con la discusión realizada en esta memoria, una posible mejora sería, fundamentalmente, aumentar y balancear el número de muestras de la base de datos. Aparte de esta mejora, podría abordarse también la aplicación de diferentes análisis o replantear desde otra perspectiva los ya utilizados, el proceso de extracción de características y/o parametrización o, incluso, una mejora de la propia herramienta utilizada para la recolección de muestras de habla inducida y espontánea: el *software* Prognosis. Este *software* podría ser perfeccionado en aras de conseguir grabaciones más naturales de los sujetos, por ejemplo, mediante una selección más personalizada del material mostrado a los participantes durante las entrevistas, actualmente adaptado a exclusivamente a la edad, o mediante la selección de diferentes tipos de contenido como es el musical. Otra línea de mejora, aunque más a largo plazo, podría centrarse una evaluación integrada en el propio *software* que permitiera valorar objetivamente la presencia o ausencia de la enfermedad o, incluso, cuantificar el grado de deterioro del lenguaje y, por ende, su relación con la AD.

Respecto al proceso de parametrización, hay que decir que, aparte de las variables temporales y de Temperatura Emocional aquí utilizadas, podría ser interesante añadir otras variables relacionadas con la calidad de la voz, variables de tipo frecuencial o *cepstral* así como características lingüísticas, bien relacionadas con el contenido y uso del lenguaje e, incluso, llegar al análisis de información paralingüística.

Aunque de los métodos automáticos localizados en esta tesis hasta la fecha no han sido lo suficientemente sofisticados para alcanzar o mejorar los resultados del entrevistador humano, en cualquier caso, la automatización eficiente del proceso de entrevistas para la detección precoz y/o control evolutivo de la AD se presenta como un paso más en el desarrollo de soluciones eHealth 4.0 basadas en voz. También supondría un paso más en la democratización del control evolutivo al poder administrarse de una forma más fácil, rápida, no invasiva y escalable. Proporcionaría aún más parámetros objetivos que los métodos tradicionales y facilitaría sustancialmente el trabajo del médico especialista. Automatizar las entrevistas no requería de ninguna infraestructura extensa, adicional a las utilizadas en el proceso manual de entrevistas y podrían usarse, si fuera necesario, incluso de forma remota como una solución *Telecare*.

El análisis multimodal podría ser otro medio para evaluar el inicio temprano de la AD mediante el análisis de patrones de comportamiento. Aparte de la voz o habla tratada en esta tesis, algunos de los rasgos biométricos conductuales que podrían incluirse en un posible análisis multimodal son la escritura o el dibujo, entre otros. Ampliando el marco concreto de este estudio, como ya se ha indicado, sería deseable un diagnóstico diferencial de la señal de voz o del habla y, a partir de él, las evaluaciones multimodales podrían complementar el diagnóstico.

A partir de medidas acústicas específicas utilizadas para diferenciar, por ejemplo, entre la enfermedad de Alzheimer y la enfermedad de Parkinson, también sería necesario determinar la naturaleza del problema, si se debe a una disfunción articulatoria, cognitiva, fonatoria o es causada por las cuerdas vocales, entre otras. Esto significa que, aunque para cada patología hay características adecuadas para medir el problema específico, es necesario identificar qué características describen qué desorden específico. Aun así, sobre la AD, hay numerosos estudios en curso cuya finalidad es arrojar luz sobre este punto y la caracterización diferencial con respecto a otras patologías neurodegenerativas, como pueden ser la enfermedad de Parkinson o también la Esclerosis Lateral Amiotrófica (ELA).

La realidad actual que vivimos se encuentra inmersa en un auténtico cambio de paradigma en todos los ámbitos de la vida, desencadenado principalmente por el avance de la tecnología. En un contexto como este, cada vez más, toman fuerza conceptos como el *eHealth*, la Telemedicina o la medicina digital. En esta tesis hemos querido comprobar objetivamente el rendimiento de un sistema conversacional automatizado o entrevistador automático concreto aplicado a la detección de la enfermedad de Alzheimer a partir del habla. Si bien se han detectado ciertas limitaciones con respecto a su homólogo manual, es necesario recalcar que, tal y como se ha demostrado en determinados casos, estos resultados son muy cercanos a los obtenidos mediante el método tradicional de entrevistas y son, también, especialmente prometedores si se tienen en cuenta los puntos de mejora propuestos. Considerando lo reflejado en esta memoria una mera prueba de concepto, los resultados obtenidos invitan sin duda a continuar profundizando y desarrollando sistemas conversacionales automatizados que permitan tomar muestras de voz de pacientes de Alzheimer con las ventajas aportadas por un método automático, escalable, objetivo, de bajo coste, no invasivo y, además, aplicable remotamente.

Referencias

[1] J. L. M. Guix, “Papel de los biomarcadores en el diagnóstico precoz de la enfermedad de Alzheimer,” *Rev. Esp. Geriatr. Gerontol.*, vol. 46, pp. 39–41, 2011.

[2] A. V Valle and A. D. L. Cortés, “Correlatos neuroanatómicos y el déficit lingüístico en la enfermedad de alzheimer: diagnóstico temprano,” *Rev. Grafías*, vol. 28, pp. 21–38, 2015.

[3] Alzheimer’s Association, “2017 Alzheimer’s disease facts and figures,” *Alzheimer’s Dement.*, vol. 13, no. 4, pp. 325–373, 2017.

[4] L. Escobar and N. P. AFANADOR, “Calidad de vida del cuidador familiar y dependencia del paciente con Alzheimer,” *Av. en Enfermería*, vol. 28, no. 1, pp. 116–128, 2010.

[5] Alzheimer’s Association, “2016 Alzheimer’s disease facts and figures,” *Alzheimer’s Dement.*, vol. 12, no. 4, pp. 459–509, 2016.

[6] M. Prince, R. Bryce, E. Albanese, A. Wimo, W. Ribeiro, and C. P. Ferri, “The global prevalence of dementia : A systematic review and metaanalysis,” *Alzheimer’s Dement.*, vol. 9, no. 1, pp. 63-75.e2, 2013.

[7] S. Seshadri, P. Wolf, and et al., “Lifetime risk of dementia and Alzheimer’s disease the impact of mortality on risk estimates in the Framingham study,” *Neurology*, vol. 49, no. 6, pp. 1498–1504, 1997.

[8] L. E. Hebert, P. A. Scherr, J. J. McCann, L. A. Beckett, and D. A. Evans, “Is the Risk of Developing Alzheimer’s Disease Greater for Women than for Men?,” *Hebert, L. E., Scherr, P. A., McCann, J. J., Beckett, L. A., Evans, D. A. Am. J. Epidemiol.*, vol. 153, no. 2, pp. 132–136, 2001.

[9] C. Ferri, M. Prince, C. Brayne, H. Brodaty, L. Fratiglioni, and M. Ganguli, “Global prevalence of dementia: a Delphi consensus study,” *Lancet*, vol. 366, no. 9503, pp. 2112–7, 2005.

[10] M. D. Hurd, “Monetary Costs of Dementia in the United States,” *N. Engl. J. Med.*, vol. 368, no. 14, pp. 1326–1334, 2013.

[11] B. Plassman, K. Langa, and et al., “Prevalence of dementia in the United States: The Aging, Demographics, and Memory Study,” *Neuroepidemiology*, vol. 29, no. 1–2, pp. 125–32, 2007.

[12] E. M. Friedman, R. A. Shih, K. M. Langa, and M. D. Hurd, “US Prevalence And

Predictors Of Informal Caregiving For Dementia," *Health Aff.*, vol. 34, no. 10, pp. 1637–41, 2015.

[13] B. Winblad, P. Amouyel, S. Andrieu, and C. Ballard, "Defeating Alzheimer's disease and other dementias: a priority for European science and society," *Lancet Neurol.*, vol. 15, no. 5, pp. 455–532, 2016.

[14] J. M. Atance, A. I. Yusta, and B. G. Grupeli, "Costs study in Alzheimer's disease," *Rev. Clínica Española*, vol. 204, no. 2, pp. 64–69, 2004.

[15] A. Gustavsson *et al.*, "Predictors of costs of care in Alzheimer ' s disease: A multinational sample of 1222 patients," *Alzheimer's Dement.*, vol. 7, pp. 318–327, 2011.

[16] R. Schulz and S. R. Beach, "Caregiving as a Risk Factor for Mortality The Caregiver Health Effects Study," *Jama*, vol. 282, no. 23, pp. 2215–2219, 1999.

[17] V. A. Freedman and B. C. Spillman, "Disability and Care Needs of Older Americans by Dementia Status: An Analysis of the 2011 National Health and Aging Trends Study. U.S.," *Office of the Assistant Secretary for Planning and Evaluation*, 2014. [Online]. Available: https://aspe.hhs.gov/report/disability-and-care-needs-older-americans-analysis-2011-national-health-and-aging-trends-study.

[18] "Caregiving in the U.S.," *National Alliance for Caregiving and AARP.*, 2015. [Online]. Available: http://www.caregiving.org/wp-content/uploads/, (2015). Available at:%0AAccessed, 2015/05/2015_CaregivingintheUS_Final-Report-June-4_WEB.pdf.

[19] M. Teresa, J. Ignacio, M. T. Algado, Á. Basterra, and J. Ignacio, "Familia y enfermedad de alzheimer. Una perspectiva cualitativa," *An. Psicol.*, vol. 13, no. 1, pp. 19–29, 1997.

[20] K. Lorenz, P. P. Freddolino, A. Comas-herrera, M. Knapp, and J. Damant, "Technology-based tools and services for people with dementia and carers : Mapping technology onto the dementia care pathway," *Dementia*, p. 1471301217691617, 2017.

[21] L. Boots, M. Vugt, R. Knippenberg, G. Kempen, and F. Verhey, "A systematic review of Internet-based supportive interventions for caregivers of patients with dementia," *Int. J. Geriatr. Psychiatry*, vol. 29, no. 4, pp. 331–344, 2014.

[22] S. J. Czaja, D. Loewenstein, R. Schulz, S. N. Nair, and D. Perdomo, "A Videophone Psychosocial Intervention for Dementia Caregivers," *Am. J. Geriatr. Psychiatry*, vol. 21, no. 13, pp. 1071–1081, 2013.

[23] E. Stanmore, B. Stubbs, D. Vancampfort, E. D. De Bruin, and J. Firth, "Neuroscience and Biobehavioral Reviews The e ff ect of active video games on cognitive functioning in clinical and non-clinical populations: A meta-analysis of randomized controlled trials," *Neurosci. Biobehav. Rev.*, vol. 78, no. April, pp. 34–43, 2017.

[24] J. Powell, T. Chiu, and G. Eysenbach, "A systematic review of networked technologies supporting carers of people with dementia," *J. Telemed. Telecare*, vol. 14, no. 3, pp. 154–156, 2008.

[25] G. Lancioni *et al.*, "A technology-aided program for helping persons with Alzheimer's disease perform daily activities," *J. Enabling Technol.*, vol. 11, no. 3, pp. 85–91, 2017.

[26] G. JE, M. Reese, and R. Tanler, "Care to Plan: An online tool that offers tailored support to dementia caregivers," *Gerontologist*, vol. 56, no. 6, pp. 1161–74, 2016.

[27] L. D. Van Mierlo, F. J. M. Meiland, H. P. J. Van Hout, and R. M. Dröes,

"Toward an evidence-based implementationmodel and checklist for personalized dementia care in the community," *Int. psychogeriatrics*, vol. 28, no. 5, pp. 801–13, 2016.

[28] Z. Ally *et al.*, "The Savvy Caregiver Program: Impact of an evidence-based intervention on the well-being of ethnically diverse caregivers," *J. Gerontol. Soc. Work*, vol. 57, no. 6–7, pp. 681–693, 2014.

[29] C. Laske, H. R. Sohrabi, S. M. Frost, K. López-de-Ipiña, P. Garrard, and M. Buscema, "Innovative diagnostic tools for early detection of Alzheimer's disease," *Alzheimer's Dement.*, vol. 11, no. 5, pp. 561–578, 2015.

[30] G8 Health and Science Ministers, "G8 Dementia Summit Declaration," *University of Toronto Library and the G7 Research Group at the University of Toronto*, 2013. [Online]. Available: https://www.gov.uk/government/publications/g8-dementiasummit-%0Aagreements/g8-dementia-summit-declaration. [Accessed: 28-Aug-2017].

[31] OECD, "Global action to drive innovation in Alzheimer's disease and other dementias: connecting research, regulation and access," in *2nd Lausanne Workshop*, 2015.

[32] M. Kivipelto, A. Solomon, and S. Ahtiluoto, "The Finnish Geriatric Intervention Study to Prevent Cognitive Impairment and Disability (FINGER): study design and progress," *Alzheimers Dement*, vol. 9, no. 6, pp. 657–65, 2013.

[33] et al. Carrié I, van Kan GA, Gillette-Guyonnet S, "Recruitment strategies for preventive trials. The MAPT study (MultiDomain Alzheimer Preventive Trial)," *J. Nutr. Health Aging*, vol. 16, no. 4, pp. 355–359, 2012.

[34] et al. Richard E, Van den Heuvel E, Moll van Charante EP, "Prevention of dementia by intensive vascular care (PreDIVA): a cluster-randomized trial in progress," *Alzheimer's Dissease Assoc. Disord.*, vol. 23, no. 3, pp. 198–204, 2009.

[35] E. Richard, S. Jongstra, and H. Soininen, "Healthy Ageing Through Internet Counselling in the Elderly: Project," *HATICE. BMJ Publ. Gr. Ltd.*, vol. 6, no. 6, 2015.

[36] M. L. Barragán-Pulido, J. B. Alonso-Hernández, Ferrer-Ballester; M. A., C. M. Travieso-González, J. Mekyska, and Z. Smékal, "Alzheimer's disease and automatic speech analysis: a review," *Expert Syst. Appl.*, vol. 150, p. 113213, 2020.

[37] C. Patterson, J. W. Feightner, A. Garcia, G. Y. R. Hsiung, C. MacKnight, and A. D. Sadovnick, "Diagnosis and treatment of dementia: 1. Risk assessment and primary prevention of Alzheimer disease," *Can. Med. Assoc. J.*, vol. 178, no. 5, pp. 548–556, 2008.

[38] P. D. McKhann G, Drachman D, Folstein M, Katzman R and S. EM., "Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's disease," *Neurology*, vol. 34, no. 7, pp. 939–44., 1984.

[39] Z. Khachaturian, "Diagnosis of Alzheimer's disease," *Arch. Neurol.*, vol. 42, no. 11, pp. 1097–105, 1985.

[40] P. J. Snyder *et al.*, "Assessment of cognition in mild cognitive impairment: A comparative study," *Alzheimer's Dement.*, vol. 7, no. 3, pp. 338–355, 2011.

[41] G. McKhann *et al.*, "The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's Dement.*, vol. 7, no. 3, pp. 263–269, 2011.

[42] M. S. Albert *et al.*, "The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's Dement.*, vol. 7, no. 3, pp. 270–279, 2011.

[43] R. A. Sperling *et al.*, "Toward defining the preclinical stages of Alzheimer ' s disease : Recommendations from the National Institute on Aging-Alzheimer ' s Association workgroups on diagnostic guidelines for Alzheimer ' s disease," *Alzheimer's Dement.*, vol. 7, no. 3, pp. 280–292, 2011.

[44] B. Dubois *et al.*, "Revising the definition of Alzheimer's disease: A new lexicon," *Lancet Neurol.*, vol. 9, no. 11, pp. 1118–1127, 2010.

[45] A. Connolly, E. Gaehl, H. Martin, J. Morris, and N. Purandare, "Underdiagnosis of dementia in primary care: variations in the observed prevalence and comparisons to the expected prevalence.," *Aging Ment. Health*, vol. 15, no. 8, pp. 978–984, 2011.

[46] R. J. Bateman *et al.*, "Clinical and biomarker changes in dominantly inherited Alzheimer's disease.," *N. Engl. J. Med.*, vol. 367, no. 9, pp. 795–804, 2012.

[47] F. T. Hane, M. Robinson, B. Y. Lee, O. Bai, and Z. Leonenko, "Recent Progress in Alzheimer's Disease Research, Part 3 : Diagnosis and Treatment," *J. Alzheimer's Dis.*, vol. 57, no. 3, pp. 645–665, 2017.

[48] C. R. Jack Jr *et al.*, "Brain beta-amyloid measures and magnetic resonance imaging atrophy both predict time-to-progression from mild cognitive impairment to Alzheimer's disease," *Brain*, vol. 133, no. 11, pp. 3336–3348., 2010.

[49] C. R. Jack *et al.*, "Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade," *Lancet Neurol.*, vol. 9, no. 1, pp. 119–128, 2010.

[50] O. L. Lopez *et al.*, "Predicting cognitive decline in Alzheimer's disease: an integrated analysis," *Alzheimer's Dement.*, vol. 6, no. 6, pp. 431–439, 2010.

[51] R. C. Petersen *et al.*, "Vitamin E and donepezil for the treatment of mild cognitive impairment," *N. Engl. J. Med.*, vol. 352, no. 23, pp. 2379–2388, 2005.

[52] C. R. Jack Jr, R. C. Petersen, M. Grundman, S. Jin, A. Gamst, and C. P. Ward, "Longitudinal MRI findings from the vitamin E and donepezil treatment study for MCI," *Neurobiol. aging*, vol. 29, no. 9, pp. 1285–95, 2008.

[53] C. C. Cummings JL, Doody R, "Disease-modifying therapies for Alzheimer disease: challenges to early intervention," *Neurology*, vol. 69, no. 16, pp. 1622–34, 2007.

[54] H. Cai, G. Li, S. Hua, Y. Liu, and L. Chen, "Effect of exercise on cognitive function in chronic disease patients : a meta-analysis and systematic review of randomized controlled trials," *Clin. Interv. Aging*, vol. 12, pp. 773–783, 2017.

[55] L. M. Gonnerman, J. M. Aronoff, A. Almor, D. Kempler, and E. S. Andersen, "From Beetle to Bug: Progression of Error Types in Naming in Alzheimer's Disease," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2004, vol. 26, p. 26.

[56] V. Taler and N. A. Phillips, "Language performance in Alzheimer's disease and mild cognitive impairment: a comparative review," *J. Clin. Exp. Neuropsychol.*, vol. 30, no. 5, pp. 501–556, 2008.

[57] P. Piscopo, G. Tosto, C. Belli, and G. Talarico, "SORL1 gene is associated with the conversion from mild cognitive impairment to Alzheimer's disease," *J. Alzheimer's Dis.*, vol. 46, no. 3, pp. 771–776, 2015.

[58] J. Matias-Guiu and R. García-Ramos, "Primary progressive aphasia: from syndrome to disease," *Neurol. (English Ed.*, vol. 28, no. 6, pp. 366–374, 2013.

[59] B. Mirheidari, D. Blackburn, and et al., "Toward the Automation of Diagnostic Conversation Analysis in Patients with Memory Complaints," *J. Alzheimer's Dis.*, vol. 58, no. 2, pp. 373–387, 2017.
[60] & S. Weintraub, S., Wicklund, A. H., "The neuropsychological profile of Alzheimer disease," *Cold Spring Harb. Perspect. Med.*, vol. 2, no. 4, 2012.
[61] G. Szatloczki, "Speaking in Alzheimer ' s disease , is that an early sign ? importance of changes in language abilities in Alzheimer ' s disease," *Front. Aging Neurosci.*, vol. 7, no. October, pp. 1–7, 2015.
[62] V. N. López, "Evaluación del discurso de las personas con Enfermedad de Alzheimer: una revisión." 2016.
[63] J. J. . Meilan, F. Martinez-Sanchez, J. Carro, N. Carcavilla, and O. Ivanova, "Voice Markers of Lexical Access in Mild Cognitive Impairment and Alzheimer's Disease," *Curr. Alzheimer Res.*, vol. 15, no. 2, pp. 111–119, 2018.
[64] R. D. Nebes, C. B. Brady, and F. J. Huff, "Automatic and attentional mechanisms of semantic priming in Alzheimer's disease," *J. Clin. Exp. Neuropsychol.*, vol. 11, no. 2, pp. 219–230, 1989.
[65] L. Bäckman, S. Jones, A.-K. Berger, E. Laukka J, and B. J. Small, "Cognitive impairment in preclinical Alzheimer's disease: a meta-analysis," *Neuropsychology*, vol. 19, no. 4, pp. 520–31, 2005.
[66] V. Deramecourt *et al.*, "Prediction of pathology in primary progressive language and speech disorders," *Neurology*, vol. 74, no. 1, pp. 42–9, 2010.
[67] E. L. Lenguaje, E. N. La, and E. D. E. Alzheimer, "El lenguaje en la enfermedad de alzheimer," *Rev. Logop. Foniatría y Audiol.*, vol. 8, no. 4, pp. 199–205, 1988.
[68] T. Sjogren, H. Sjogren, and A. G. Lindgren, "Morbus Alzheimer and morbus Pick; a genetic, clinical and patho-anatomical study.," *Acta Psychiatr. Neurol. Scand. Suppl.*, vol. 82, pp. 1–152, 1952.
[69] R. S. Allison, "The Senile Brain: A Clinical Study," *Posgraduate Med. J.*, p. 656, 1962.
[70] K. Forbes-McKay, M. F. Shanks, and A. Venneri, "Profiling spontaneous speech decline in Alzheimer's disease: a longitudinal study," *Acta Neuropsychiatr.*, vol. 25, no. 06, pp. 320–327, Dec. 2013.
[71] N. H. Frijda, "Emotion, cognitive structure, and action tendency," *Cogn. Emot.*, vol. 1, no. 2, pp. 115–143, Apr. 1987.
[72] S. Spence, "Descartes' error: Emotion, reason and the human brain," *BMJ*, vol. 310, no. 6988, p. 1213, 1995.
[73] P. Maresova, M. Valis, J. Hort, and K. Kuca, "Alzheimer's disease and language impairments: social intervention and medical treatment," *Clin. Interv. Aging*, vol. 10, pp. 1401–1408, 2015.
[74] K. E. Forbes-McKay and A. Venneri, "Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task," *Neurol. Sci.*, vol. 26, no. 4, pp. 243–254, Oct. 2005.
[75] A. De Renzi, L. V.-B. a journal of Neurology, and U. 1962, "Token test: A sensitive test to detect receptive disturbances in aphasics.," *Brain A J. Neurol.*, vol. 85, pp. 665–678, 1962.
[76] H. Goodglass and E. Kaplan, "Boston naming test," 1983.
[77] F. Cuetos-Vega, M. Menéndez-González, and T. Calatayud-Noguera, "Descripción de un nuevo test para la detección precoz de la enfermedad de Alzheimer," *REV NEUROL*, vol. 44, no. 8, pp. 469–474, 2007.
[78] C. D. Hoffmann, I., Nemeth, D., Dye and J. Pákáski, M., Irinyi, T., & Kálmán, "Temporal parameters of spontaneous speech in Alzheimer's disease," *Int. J.*

Speech. Lang. Pathol., vol. 12, no. 1, pp. 29–34, 2010.

[79] R. Barbarotto, E. Capitani, T. Jori, M. Laiacona, and S. Molinari, "Picture naming and progression of Alzheimer's disease: an analysis of error types," *Neuropsychologia*, vol. 36, no. 5, pp. 397–405, May 1998.

[80] K. Wesnes, "Assessing cognitive function in clinical trials: latest developments and future directions," *Drug Discov. Today*, vol. 7, no. 1, pp. 29–35, Jan. 2002.

[81] L. Brabenec, J. Mekyska, Z. Galaz, and I. Rektorova, "Speech disorders in Parkinson's disease: early diagnostics and effects of medication and brain stimulation," *J. Neural Transm.*, vol. 124, no. 3, pp. 303–334, Mar. 2017.

[82] I. C. solutions IBM Watson, "SimpleC | IBM," 2016. [Online]. Available: https://www.ibm.com/case-studies/w796019n50088s93. [Accessed: 15-May-2019].

[83] Cambridge Cognition, "CANTAB Mobile | Cambridge Cognition," 2019. [Online]. Available: https://www.cambridgecognition.com/products/digital-healthcare-technology/cantab-mobile/. [Accessed: 15-May-2019].

[84] "Winterlight Labs," 2018. [Online]. Available: https://winterlightlabs.com/. [Accessed: 15-May-2019].

[85] ki-elements, "Δelta - Digital neurocognitive testing," 2018. [Online]. Available: https://ki-elements.de/. [Accessed: 15-May-2019].

[86] Veritone, "Voice Analysis to Detect Alzheimer's Disease within Seconds," 2019. [Online]. Available: https://www.veritone.com/blog/voice-analysis-detects-alzheimers-disease/. [Accessed: 15-May-2019].

[87] J. J. G. Meilán, F. Martínez-Sánchez, J. Carro, D. E. López, L. Millian-Morell, and J. M. Arana, "Speech in Alzheimer's disease: can temporal and acoustic parameters discriminate dementia?," *Dement. Geriatr. Cogn. Disord.*, vol. 37, no. 5–6, pp. 327–34, 2014.

[88] K. López-de-Ipiña *et al.*, "New Approaches for Alzheimer's Disease Diagnosis Based on Automatic Spontaneous Speech Analysis and Emotional Temperature," in *Ambient Assisted Living and Home Care. IWAAL 2012. Lecture Notes in Computer Science, vol 7657. Springer, Berlin, Heidelberg*, Springer, Berlin, Heidelberg, 2012, pp. 407–414.

[89] K. López-de-Ipiña *et al.*, "Feature selection for spontaneous speech analysis to aid in Alzheimer's disease diagnosis: A fractal dimension approach," *Comput. Speech Lang.*, vol. 30, no. 1, pp. 43–60, Mar. 2015.

[90] K. López-de-Ipiña *et al.*, "On Automatic Diagnosis of Alzheimer's Disease Based on Spontaneous Speech Analysis and Emotional Temperature," *Cognit. Comput.*, vol. 7, no. 1, pp. 44–55, Feb. 2015.

[91] P. Gómez-Vilda, O. P.-B.-… W. C. On, and U. 2014, "Biomechanical characterization of phonation in Alzheimer's Disease," in *3rd IEEE International Work-Conference on Bioinspired Intelligence*, 2014.

[92] J. B. Alonso, J. Cabrera, M. Medina, and C. M. Travieso, "New approach in quantification of emotional intensity from the speech signal: emotional temperature," *Expert Syst. Appl.*, vol. 42, no. 24, pp. 9554–9564, 2015.

[93] V. Baldas, C. Lampiris, C. Capsalis, and D. Koutsouris, "Early Diagnosis of Alzheimer's Type Dementia Using Continuous Speech Recognition," in *Wireless Mobile Communication and Healthcare. MobiHealth 2010. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 55*, Springer, Berlin, Heidelberg, 2011, pp. 105–110.

[94] J. J. G. Meilán, F. Martínez-Sánchez, J. Carro, J. A. Sánchez, and E. Pérez, "Acoustic Markers Associated with Impairment in Language Processing in

Alzheimer's Disease," *Span. J. Psychol.*, vol. 15, no. 02, pp. 487–494, Jul. 2012.

[95] A. Satt *et al.*, "Evaluation of Speech-Based Protocol for Detection of Early-Stage Dementia," in *INTERSPEECH 2013*, 2013, pp. 1692–1696.

[96] A. Satt, R. Hoory, A. König, P. Aalten, and P. H. Robert, "Speech-Based Automatic and Robust Detection of Very Early Dementia," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014, pp. 2538–2542.

[97] A. Khodabakhsh and C. Demiroglu, "Analysis of Speech-Based Measures for Detecting and Monitoring Alzheimer's Disease," in *Data Mining in Clinical Medicine*, Humana Press, New York, NY, 2015, pp. 159–173.

[98] M. Nasrolahzadeh, Z. Mohammadpoory, and J. Haddadnia, "A novel method for early diagnosis of Alzheimer's disease based on higher-order spectral estimation of spontaneous speech signals," *Cogn. Neurodyn.*, vol. 10, no. 6, pp. 495–503, Dec. 2016.

[99] H. Tanaka, H. Adachi, N. Ukita, T. Kudo, and S. Nakamura, "Automatic detection of very early stage of dementia through multimodal interaction with computer avatars," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016*, 2016, pp. 261–265.

[100] H. Tanaka, H. Adachi, N. Ukita, T. Kudo, and S. Nakamura, "Automatic Detection of Very Early Stage of Dementia through Spoken Dialog with Computer Avatars," *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. pp. 261–265, 2016.

[101] D. Beltrami *et al.*, "Automatic Identification of Mild Cognitive Impairment through the Analysis of Italian Spontaneous Speech Productions," in *LREC*, 2016, pp. 2089–2093.

[102] A. König, A. Satt, R. David, and P. Robert, "Innovative voice analytics for the assessment and monitoring of cognitive decline in people with dementia and mild cognitive impairment," *Alzheimer's Dement.*, vol. 12, no. 7, p. P363, Jul. 2016.

[103] J. Tröger, N. Linz, J. Alexandersson, A. König, and P. H. Robert, "Automated speech-based screening for alzheimer's disease in a care service scenario," in *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare, ACM*, 2017, pp. 292–297.

[104] S. Mirzaei, M. El Yacoubi, S. Garcia-Salicetti, J. Boudy, C. K. S. Muvingi, and V. Cristancho-Lacroix, "Automatic speech analysis for early Alzeimer's disease diagnosis," in *JETSAN 2017: 6e Journées d'Etudes sur la Télésant*, 2017, pp. 114–116.

[105] V. Rentoumi *et al.*, "Automatic detection of linguistic indicators as a means of early detection of Alzheimer's disease and of related dementias: A computational linguistics analysis," in *Cognitive Infocommunications (CogInfoCom), 8th IEEE International Conference*, 2017, pp. 33–38.

[106] F. Martínez-Sánchez *et al.*, "Speech rhythm alterations in Spanish-speaking individuals with Alzheimer's disease," *Aging, Neuropsychol. Cogn.*, vol. 24, no. 4, pp. 418–434, Jul. 2017.

[107] S. Kato, A. Homma, and T. Sakuma, "Easy Screening for Mild Alzheimer's Disease and Mild Cognitive Impairment from Elderly Speech," *Curr. Alzheimer Res.*, vol. 15, no. 2, pp. 104–110, 2018.

[108] L. Toth, I. Hoffmann, G. Gosztolya, V. Vincze, G. Szatloczki, and Z. Banreti, "A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech," *Curr. Alzheimer Res.*, vol. 15, no. 2, pp. 130–138, 2018.

[109] A. Satt, R. Hoory, A. König, P. Aalten, and P. H. Robert, "Speech-based automatic and robust detection of very early dementia," in *Fifteenth Annual Conference of the International Speech Communication Association. INTERSPEECH-2014*, 2014, pp. 2538–2542.

[110] A. König *et al.*, "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease," *Alzheimer's Dement. Diagnosis, Assess. Dis. Monit.*, vol. 1, no. 1, pp. 112–124, Mar. 2015.

[111] S. Al-Hameed, M. Benaissa, and H. Christensen, "Simple and robust audio-based detection of biomarkers for Alzheimer's disease," in *7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2016, pp. 32–36.

[112] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data mining: practical machine learning tools and techniques*. 2016.

[113] T. Warnita, N. Inoue, and K. Shinoda, "Detecting Alzheimer's Disease Using Gated Convolutional Neural Network from Audio Data," *arXiv preprint arXiv:1803.11344.*, 30-Mar-2018. [Online]. Available: http://arxiv.org/abs/1803.11344. [Accessed: 01-May-2018].

[114] K. D. Mueller, R. L. Koscik, B. P. Hermann, S. C. Johnson, and L. S. Turkstra, "Declines in Connected Language Are Associated with Very Early Mild Cognitive Impairment: Results from the Wisconsin Registry for Alzheimer's Prevention," *Front. Aging Neurosci.*, vol. 9, p. 437, Jan. 2018.

[115] "Proyectos Fundación CITA Alzheimer," 2017. [Online]. Available: http://www.cita-alzheimer.org/investigacion/proyectos. [Accessed: 26-May-2018].

[116] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Commun.*, vol. 40, no. 1–2, pp. 5–32, Apr. 2003.

[117] Y. D. Chavhan, B. S. Yelure, and K. N. Tayade, "Speech emotion recognition using RBF kernel of LIBSVM," in *2015 2nd International Conference on Electronics and Communication Systems (ICECS)*, 2015, pp. 1132–1135.

[118] H. Balti and A. S. Elmaghraby, "Emotion analysis from speech using temporal contextual trajectories," in *2014 IEEE Symposium on Computers and Communications (ISCC)*, 2014, pp. 1–7.

[119] P. Laukka, *Vocal expression of emotion : discrete-emotions and dimensional accounts*. Acta Universitatis Upsaliensis, Uppsala universitet, 2004.

[120] M. Goudbeek and K. Scherer, "Beyond arousal: Valence and potency/control cues in the vocal expression of emotion," *J. Acoust. Soc. Am.*, vol. 128, no. 3, p. 1322, Sep. 2010.

[121] O. Kwon, K. Chan, and et al., "Emotion recognition by speech signals," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[122] C. Lee and S. Narayanan, "Emotion recognition using a data-driven fuzzy inference system," in *Eighth European conference on speech communication and technology*, 2003.

[123] A. Harimi, A. Shahzadi, and A. Ahmadyfard, "Recognition of emotion using non-linear dynamics of speech," in *7'th International Symposium on Telecommunications (IST'2014)*, 2014, pp. 446–451.

[124] H. Altun and G. Polat, "Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 8197–8203, May 2009.

[125] K. Lopez-de-Ipiña *et al.*, "Alzheimer disease diagnosis based on automatic spontaneous speech analysis," in *International Joint Conference on Computational Intelligence. "IJCCI 2012: proceedings of the 4th International*

Joint Conference on Computational Intelligence: Barcelona, Spain: 5-7 October, 2012," 2012, pp. 698–705.
[126] P. Barrett, "Voice activity detector," US005749067A, 1998.
[127] T. M. Rath and R. Manmatha, "Word Spotting for Historical Documents," *Int. J. Doc. Anal. Recognit.*, vol. 9, no. 2, pp. 139–152, 2007.
[128] J. Weiner and T. Schultz, "Detection of intra-personal development of cognitive impairment from conversational speech," in *Speech Communication; 12th ITG Symposium; Proceedings of VDE.*, 2016, pp. 1–5.
[129] J. Weiner, C. Herff, and T. Schultz, "Speech-Based Detection of Alzheimer's Disease in Conversational German.," in *INTERSPEECH*, 2016, pp. 1938–1942.
[130] R. Sadeghian, J. Schaffer, and S. A. Zahorian, "Speech Processing Approach for Diagnosing Dementia in an Early Stage," in *Proceedings Interspeech 2017*, 2017, pp. 2705–2709.
[131] L. Hernández-Domíngue, E. García-Canó, S. Ratt, and G. Sierra-Martínez, "Detection of Alzheimer's disease based on automatic analysis of common objects descriptions," in *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, 2016, pp. 10–15.
[132] V. Vincze *et al.*, "Detecting Mild Cognitive Impairment by Exploiting Linguistic Information from Transcripts," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 181–187.
[133] C. Thomas, V. Keselj, N. Cercone, K. Rockwood, and E. Asp, "Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech," in *IEEE International Conference Mechatronics and Automation, 2005*, 2005, vol. 3, pp. 1569–1574.
[134] L. Tóth, G. Gosztolya, V. Vincze, I. Hoffmann, G. Szatlóczki, and E. Biró, "Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech Using ASR," in *INTERSPEECH-2015*, 2015, pp. 2694–2698.
[135] S. Wankerl, E. Nöth, and S. Evert, "An N-gram based approach to the automatic diagnosis of Alzheimer's disease from spoken language," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.*, 2017.
[136] B. Mirheidari, D. Blackburn, and et al., "Diagnosing people with dementia using automatic conversation analysis," in *Proceedings of Interspeech, ISCA.*, 2016, pp. 1220–1224.
[137] L. Hernández-Domínguez, S. Ratté, G. Sierra-Martínez, and A. Roche-Bergua, "Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task," *Alzheimer's Dement. Diagnosis, Assess. Dis. Monit.*, vol. 10, pp. 260–268, 2018.
[138] A. Khodabakhsh, F. Yesil, E. Guner, and C. Demiroglu, "Evaluation of linguistic and prosodic features for detection of Alzheimer's disease in Turkish conversational speech," *EURASIP J. Audio, Speech, Music Process.*, vol. 2015, no. 1, p. 9, Dec. 2015.
[139] M. Asgari, J. Kaye, and H. Dodge, "Predicting mild cognitive impairment from spontaneous spoken utterances," *Alzheimer's Dement. Transl. Res. Clin. Interv.*, vol. 3, no. 2, pp. 219–228, 2017.
[140] L. Tóth *et al.*, "Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech Using ASR," in *Sixteenth Annual Conference of the International Speech Communication Association. INTERSPEECH-2015*, 2015, pp. 2694–2698.
[141] M. Abdalla, F. Rudzicz, and G. Hirst, "Rhetorical structure and Alzheimer's

disease," *Aphasiology*, vol. 32, no. 1, pp. 41–60, Jan. 2018.
[142] W. C. Mann and S. A. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization," *Text-Interdisciplinary J. Study Discourse*, vol. 8, no. 3, pp. 243–281, 1988.
[143] L. Carlson and D. Marcu, "Discourse tagging reference manual (Tech. Rep.)," *Univ. South. California. Inf. Sci. Institute.*, 2001.
[144] M. Yancheva, "Automatic assessment of information content in speech for detection of dementia of the Alzheimer type," 2016.
[145] K. Sirts, O. Piguet, and M. Johnson, "Idea density for predicting Alzheimer's disease from transcribed speech," *arXiv Prepr. arXiv1706.04473.*, Jun. 2017.
[146] S. M. Aluisio, A. Cunha, C. Toledo, and C. Scarton, "A computational tool for automated language production analysis aimed at dementia diagnosis," in *International Conference on Computational Processing of the Portuguese Language, XII; Demonstration Session*, 2016.
[147] S. Aluísio, A. Cunha, and C. Scarton, "Evaluating Progression of Alzheimer's Disease by Regression and Classification Methods in a Narrative Language Test in Portuguese," in *International Conference on Computational Processing of the Portugues*, 2016, pp. 109–114.
[148] C. M. Toledo *et al.*, "Analysis of macrolinguistic aspects of narratives from individuals with Alzheimer's disease, mild cognitive impairment, and no cognitive impairment," *Alzheimer's Dement. Diagnosis, Assess. Dis. Monit.*, vol. 10, pp. 31–40, 2018.
[149] S. C. Johnson *et al.*, "The Wisconsin Registry for Alzheimer's Prevention: A review of findings and current directions," *Alzheimer's Dement. Diagnosis, Assess. Dis. Monit.*, vol. 10, pp. 130–142, 2018.
[150] B. Macwhinney, D. Fromm, M. Forbes, and A. Holland, "AphasiaBank: Methods for Studying Discourse.," *Aphasiology*, vol. 25, no. 11, pp. 1286–1307, 2011.
[151] C. Pye and B. MacWhinney, "The CHILDES Project: Tools for Analyzing Talk," *Language (Baltim).*, vol. 70, no. 1, p. 156, Mar. 1994.
[152] Jiri Mekyska *et al.*, "Robust and complex approach of pathological speech signal analysis," *Neurocomputing*, vol. 167, pp. 94–111, 2015.
[153] G. Vaziri, F. Almasganj, and R. Behroozmand, "Pathological assessment of patients' speech signals using nonlinear dynamical analysis," *Comput. Biol. Med.*, vol. 40, no. 1, pp. 54–63, 2010.
[154] M. Little, P. McSharry, and E. Hunter, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1015–1022, 2009.
[155] J. Shao, J. MacCallum, Y. Zhang, and A. Sprecher, "Acoustic analysis of the tremulous voice: assessing the utility of the correlation dimension and perturbation parameters," *J. Commun. Disord.*, vol. 43, no. 1, p. 35.44, 2010.
[156] R. Esteller and et al., "A comparison of waveform fractal dimension algorithms," *IEEE Trans. Circuits Syst. I Fundam. Theory Appl.*, vol. 48, no. 2, pp. 177–183, 2001.
[157] M. Aboy, R. Hornero, and D. Abásolo, "Interpretation of the Lempel-Ziv complexity measure in the context of biomedical signal analysis," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 11, pp. 2282–2288, 2006.
[158] J. Orozco-Arroyave and S. Murillo-Rendón, "Automatic selection of acoustic and non-linear dynamic features in voice signals for hypernasality detection," in *Twelfth Annual Conference of the International Speech Communication Association INTERSPEECH*, 2011, pp. 529–532.

[159] P. Henríquez, J. B. Alonso, M. A. Ferrer, C. M. Travieso, J. I. Godino-Llorente, and F. Díaz-De-María, "Characterization of Healthy and Pathological Voice Through Measures Based on Nonlinear Dynamics," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 17, no. 6, pp. 1186–1195, 2009.

[160] A. W. Jayawardena, P. Xu, and W. K. Li, "Modified correlation entropy estimation for a noisy chaotic time series," *Chaos An Interdiscip. J. Nonlinear Sci.*, vol. 20, no. 2, p. 023104, Jun. 2010.

[161] J. M. Yentes, N. Hunt, K. K. Schmid, J. P. Kaipust, D. McGrath, and N. Stergiou, "The Appropriate Use of Approximate Entropy and Sample Entropy with Short Data Sets," *Ann. Biomed. Eng.*, vol. 41, no. 2, pp. 349–365, Feb. 2013.

[162] H. Herisa Khadivi, B. Seyed Aghazadehb, and M. Nikkhah-Bahramic, "Optimal feature selection for the assessment of vocal fold disorders," *Comput. Biol. Med.*, vol. 39, no. 10, pp. 860–868, 2009.

[163] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, and I. M. Moroz, "Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection," *Biomed. Eng. Online*, vol. 6, no. 1, p. 23, 2007.

[164] A. Tsanas, … M. L.-J. of the royal, and U. 2011, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity," *J. R. Soc. Interface*, vol. 8, no. 59, pp. 842–855, 2011.

[165] M. Nasrolahzadeh, Z. Mohammadpoori, and J. Haddadnia, "Analysis of mean square error surface and its corresponding contour plots of spontaneous speech signals in Alzheimer's disease with adaptive wiener filter," *Comput. Human Behav.*, vol. 61, pp. 364–371, Aug. 2016.

[166] K. López-de-Ipiña *et al.*, "On the Selection of Non-Invasive Methods Based on Speech Analysis Oriented to Automatic Alzheimer Disease Diagnosis," *Sensors*, vol. 13, no. 5, pp. 6730–6745, May 2013.

[167] K. Lopez-de-Ipina *et al.*, "Automatic analysis of emotional response based on non-linear speech modeling oriented to Alzheimer disease diagnosis," in *2013 IEEE 17th International Conference on Intelligent Engineering Systems (INES)*, 2013, pp. 61–64.

[168] K. Lopez-de-Ipiña *et al.*, "Spontaneous Speech and Emotional Response Modeling Based on One-Class Classifier Oriented to Alzheimer Disease Diagnosis," in *XIII Mediterranean Conference on Medical and Biological Engineering and Computing*, Springer, Cham, 2014, pp. 571–574.

[169] B. Mandelbrot, *Fractal geometry of nature*. New York: WH freeman.: Times Books, 1982.

[170] Y. X. Huang, F. G. Schmitt, J.-P. Hermand, Y. Gagne, Z. M. Lu, and Y. L. Liu, "Arbitrary-order Hilbert spectral analysis for time series possessing scaling statistics: Comparison study with detrended fluctuation analysis and wavelet leaders," *Phys. Rev. E*, vol. 84, no. 1, p. 016208, Jul. 2011.

[171] K. López-de-Ipiña *et al.*, "Feature Extraction Approach Based on Fractal Dimension for Spontaneous Speech Modelling Oriented to Alzheimer Disease Diagnosis," in *International Conference on Nonlinear Speech Processing. Advances in Nonlinear Speech Processing*, Springer, Berlin, Heidelberg, 2013, pp. 144–151.

[172] K. López-de-Ipiña, M. Faundez-Zanuy, J. Solé-Casals, F. Zelarin, and P. Calvo, "Multi-class Versus One-Class Classifier in Spontaneous Speech Analysis Oriented to Alzheimer Disease Diagnosis," in *Recent Advances in Nonlinear Speech Processing*, Springer, Cham, 2016, pp. 63–72.

[173] K. López-de-Ipiña *et al.*, "Feature selection for automatic analysis of emotional response based on nonlinear speech modeling suitable for diagnosis of Alzheimer′s disease," *Neurocomputing*, vol. 150, pp. 392–401, Feb. 2015.
[174] H. E. Stanley *et al.*, "Fractal landscapes in biological systems," *Fractals*, vol. 01, no. 03, pp. 283–301, Sep. 1993.
[175] J. W. Kantelhardt, E. Koscielny-Bunde, H. H. Rego, S. Havlin, and A. Bunde, "Detecting long-range correlations with detrended fluctuation analysis," *Phys. A Stat. Mech. its Appl.*, vol. 295, no. 3–4, pp. 441–454, 2001.
[176] S. Bhaduri, R. Das, and D. Ghosh, "Non-Invasive Detection of Alzheimer's Disease-Multifractality of Emotional Speech," *J. Neurol. Neurosci.*, vol. 7, no. 2, 2016.
[177] Z. Chen, P. C. Ivanov, K. Hu, and H. E. Stanley, "Effect of nonstationarities on detrended fluctuation analysis," *Phys. Rev. E*, vol. 65, no. 4, p. 041107, Apr. 2002.
[178] D. C. González, L. Luan Ling, and F. Violaro, "Analysis of the Multifractal Nature of Speech Signals," in *Iberoamerican Congress on Pattern Recognition*, 2012, pp. 740–748.
[179] J. M. Hausdorff, P. L. Purdon, C. K. Peng, Z. Ladin, J. Y. Wei, and A. L. Goldberger, "Fractal dynamics of human gait: stability of long-range correlations in stride interval fluctuations," *J. Appl. Physiol.*, vol. 80, no. 5, pp. 1448–1457, May 1996.
[180] J. W. Kantelhardt, S. A. Zschiegner, E. Koscielny-Bunde, S. Havlin, A. Bunde, and H. E. Stanley, "Multifractal detrended fluctuation analysis of nonstationary time series," *Phys. A Stat. Mech. its Appl.*, vol. 316, no. 1–4, pp. 87–114, 2002.
[181] P. Oświęcimka, J. Kwapień, and S. Drożdż, "Wavelet versus detrended fluctuation analysis of multifractal structures," *Phys. Rev. E*, vol. 74, no. 1, p. 016103, Jul. 2006.
[182] E. Serrano and A. Figliola, "Wavelet leaders: a new method to estimate the multifractal singularity spectra," *Phys. A Stat. Mech. its Appl.*, vol. 388, no. 14, pp. 2793–2805, 2009.
[183] "Toronto emotional speech set (TESS) | TSpace Repository," 2018. [Online]. Available: https://tspace.library.utoronto.ca/handle/1807/24487. [Accessed: 26-May-2018].
[184] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference*, 2013, pp. 3687–3691.
[185] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.
[186] U. M. de Lizarduy, P. C. Salomón, P. G. Vilda, M. E. Torres, and K. L. de Ipiña, "ALZUMERIC: A decision support system for diagnosis and monitoring of cognitive impairment," *Loquens*, vol. 4, no. 1, p. 37, 2017.
[187] R. R. Picard and R. D. Cook, "Cross-Validation of Regression Models," *J. Am. Stat. Assoc.*, vol. 79, no. 387, pp. 575–583, Sep. 1984.
[188] K. López-de-Ipiña, U. Martinez-de-Lizarduy, P. M. Calvo, B. Beitia, J. García-Melero, and M. Ecay-Torres, "Analysis of Disfluencies for automatic detection of Mild Cognitive Impartment: a deep learning approach," in *Bioinspired Intelligence (IWOBI), 2017 International Conference and Workshop IEEE.*, 2017, pp. 1–4.
[189] K. Lopez-de-Ipina, U. Martinez-de-Lizarduy, P. M. Calvo, J. Mekyska, B. Beitia, and N. Barroso, "Advances on automatic speech analysis for early detection of

Alzheimer disease: a non-linear multi-task approach," *Curr. Alzheimer Res.*, vol. 15, no. 2, pp. 139–148, 2018.
[190] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the Munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.
[191] B. Schuller *et al.*, "The INTERSPEECH 2010 paralinguistic challenge," in *Proceedings INTERSPEECH 2010*, 2010, pp. 2794–2797.
[192] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge," in *Twelfth Annual Conference of the International Speech Communication Association.*, 2011.
[193] B. Schuller *et al.*, "The interspeech 2012 speaker trait challenge," in *Thirteenth Annual Conference of the International Speech Communication Association.*, 2012.
[194] Ozone, "Auricular Rage ST - Ozone Gaming." [Online]. Available: https://www.ozonegaming.com/es/product/rage-st. [Accessed: 05-Nov-2019].
[195] GoodData, "Normality Testing - Skewness and Kurtosis," 2020. [Online]. Available: https://www.gooddata.com/.
[196] F. Wilcoxon and R. Wilcox, "Some rapid approximate statistical procedures," 1964.
[197] W. H. Kruskal and W. A. Wallis, "Use of Ranks in One-Criterion Variance Analysis," *J. Am. Stat. Assoc.*, vol. 47, no. 260, pp. 583–621, 1952.
[198] S. María Pozo Abreu, "Prueba de la Mediana. Ejemplo en SPSS."
[199] L. Zhou, K. C. Fraser, and F. Rudzicz, "Speech recognition in Alzheimer's disease and in its assessment," in *INTERSPEECH 2016*, 2016, pp. 1948–1952.
[200] D. Roy and A. Pentland, "Automatic spoken affect classification and analysis," in *Automatic Face and Gesture Recognition, Proceedings of the Second International Conference*, 1996, pp. 363–367.
[201] K. Lopez-de-Ipina *et al.*, "Automatic analysis of Categorical Verbal Fluency for Mild Cognitive impartment detection: A non-linear language independent approach," in *Bioinspired Intelligence (IWOBI), 2015 4th International Work Conference IEEE.*, 2015, pp. 101–104.
[202] "AMI Corpus," 2006. [Online]. Available: http://groups.inf.ed.ac.uk/ami/corpus/. [Accessed: 26-May-2018].
[203] M. F. Folstein, L. N. Robins, and J. E. Helzer, "The mini-mental state examination," *Arch. Gen. Psychiatry*, vol. 40, no. 7, p. 812, 1983.
[204] K. Rockwood, J. E. Graham, and S. Fay, "Goal setting and attainment in Alzheimer's disease patients treated with donepezil," *J. Neurol. Neurosurg. Psychiatry*, vol. 73, no. 5, pp. 500–507, 2002.
[205] J. Weiner, C. Frankenberg, and D. Telaar, "Towards Automatic Transcription of ILSE—an Interdisciplinary Longitudinal Study of Adult Development and Aging," *LREC*, 2016.
[206] "DementiaBank | TalkBank," 2007. [Online]. Available: https://dementia.talkbank.org/access/. [Accessed: 26-May-2018].
[207] J. B. Alonso-Hernández, M. L. Barragán-Pulido, J. M. Gil-Bordón, M. Á. Ferrer-Ballester, and C. M. Travieso-González, "Using a Human Interviewer or an Automatic Interviewer in the Evaluation of Patients with AD from Speech," *Appl. Sci.*, vol. 11, no. 7, p. 3228, 2021.
[208] M. Boyé, T. M. Tran, and N. Grabar, "NLP-Oriented Contrastive Study of Linguistic Productions of Alzheimer's and Control People," in *International*

Conference on Natural Language Processing. Advances in Natural Language Processing, 2014, pp. 412–424.

[209] S. Luz, "Longitudinal Monitoring and Detection of Alzheimer's Type Dementia from Spontaneous Speech Data," in *Computer-Based Medical Systems (CBMS), 2017 IEEE 30th International Symposium on IEEE.*, 2017, pp. 45–46.

[210] "Carolinas Conversations Collection - About - Who We Are," 2008. [Online]. Available: http://carolinaconversations.musc.edu/about/who. [Accessed: 26-May-2018].

[211] H. Peraita and L. Grasso, "Corpus lingüístico de definiciones de categorías semánticas de personas mayores sanas y con la enfermedad del alzheimer," *Tech. report, Fund. BBVA.*, 2010.

[212] J. Graovac, J. Kovacevic, and G. P. Lazetic, "Machine learning-based approach to help diagnosing Alzheimer's disease through spontaneous speech analysis," in *Belgrade BioInformatics Conference 2016*, 2016, p. 111.

[213] G. Gosztolya *et al.*, "Detecting Mild Cognitive Impairment from Spontaneous Speech by Correlation-Based Phonetic Feature Selection," in *INTERSPEECH 2016*, 2016, pp. 107–111.

[214] L. Tóth, G. Gosztolya, and et al., "Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech Using ASR," in *Sixteenth Annual Conference of the International Speech Communication Association. INTERSPEECH-2015*, 2015, pp. 2694–2698.

[215] M.-C. St-Pierre, B. Ska, and R. Béland, "Lack of coherence in the narrative discourse of patients with dementia of the Alzheimer's type," *J. Multiling. Commun. Disord.*, vol. 3, no. 3, pp. 211–215, Jan. 2005.

[216] G. Malekzadeh, G. Arsalan, and M. Shahabi, "A comparative study on the use of cohesion devices by normal age persian natives and those suffering from Alzheimer's disease," *J. Med. Sci. Islam. Azad Univ. Mashhad*, vol. 5, no. 3, pp. 153–161, 2009.

[217] A. Ahangar, S. Morteza, J. Fadaki, and A. Sehhati, "The Comparison of Morpho-Syntactic Patterns Device Comprehension in Speech of Alzheimer and Normal Elderly People," *Zahedan J Res Med Sci*, 2018.

[218] E. Kaplan, H. Goodglass, and S. Weintraub, "Boston naming test," *Pro-ed.*, 2001.

[219] "Dem@Care Project - Project," 2011. [Online]. Available: http://www.demcare.eu/. [Accessed: 26-May-2018].

[220] "Corpus Linguistics and Linguistic Theory," vol. 7, no. 1, pp. 143–161, 2011.

[221] C. De Looze *et al.*, "Changes in Speech Chunking in Reading Aloud is a Marker of Mild Cognitive Impairment and Mild-to-Moderate Alzheimer's Disease," *Curr. Alzheimer Res.*, vol. 15, no. 9, pp. 828–847, Jul. 2018.

[222] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. Mcgonigle, "The Natural History of Alzheimer's Disease: Description of Study Cohort and Accuracy of Diagnosis," *Arch. Neurol.*, vol. 51, no. 6, pp. 585–594, Jun. 1994.

[223] D. Devault *et al.*, "SimSensei Kiosk : A Virtual Human Interviewer for Healthcare Decision Support," in *2014 International Conference on Autonomous Agents and Multi-Agent Systems. International Foundation for Autonomous Agents and Multiagent Systems*, 2014, no. 1, pp. 1061–1068.

[224] D. Huggins-daines *et al.*, "POCKETSPHINX : A FREE , REAL-TIME CONTINUOUS SPEECH RECOGNITION SYSTEM FOR HAND-HELD DEVICES Language Technologies Institute (dhuggins , mohitkum , archan , awb , rkm , air)@ cs . cmu . edu," *Icassp 2006*, pp. 185–188, 2006.

[225] G. Littlewort *et al.*, "The computer expression recognition toolbox (CERT)," in *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011*, 2011, pp. 298–305.
[226] M. Stone, "Specifying Generation of Referring Expressions by Example," *Work. Pap. 2003 {AAAI} Spring Symp. Nat. Lang. Gener. Spok. Writ. Dialogue*, pp. 133–140, 2003.
[227] MINDMAKERS.ORG, "Wiki - SAIBA - Mindmakers." [Online]. Available: http://mindmakers.com/projects/SAIBA. [Accessed: 02-Apr-2021].
[228] T. Bickmore, D. Schulman, and G. Shaw, "DTask and litebody: Open source, standards-based tools for building web-deployed embodied conversational agents," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2009, vol. 5773 LNAI, pp. 425–431.
[229] I. Poggi, C. Pelachaud, F. de Rosis, V. Carofiglio, and B. De Carolis, "Greta. A Believable Embodied Conversational Agent," Springer, Dordrecht, 2005, pp. 3–25.
[230] M. Schröder, "The SEMAINE API: Towards a Standards-Based Framework for Building Emotion-Oriented Systems," *Adv. Human-Computer Interact.*, vol. 2010, pp. 1–21, Jan. 2010.
[231] A. Hartholt *et al.*, "All Together Now Introducing the Virtual Human Toolkit."
[232] D. DeVault *et al.*, "SimSensei kiosk: A virtual human interviewer for healthcare decision support," in *13th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2014*, 2014, vol. 2, no. 1, pp. 1061–1068.
[233] Universidad de Las Palmas de Gran Canaria, "Memoria digital de Canarias - mdC." [Online]. Available: https://mdc.ulpgc.es/. [Accessed: 11-Dec-2019].
[234] K. Yasuda, K. Kuwabara, N. Kuwahara, S. Abe, and N. Tetsutani, "Effectiveness of personalised reminiscence photo videos for individuals with dementia," *Neuropsychol. Rehabil.*, vol. 19, no. 4, pp. 603–619, Aug. 2009.
[235] G. Gowans, R. Dye, J. Campbell, A. Astell, N. Alm, and M. Ellis, "Designing a multimedia conversation aid for reminiscence therapy in dementia care environments," in *Conference on Human Factors in Computing Systems - Proceedings*, 2004, pp. 825–836.
[236] B. H. Davis and D. Shenk, "Beyond reminiscence: using generic video to elicit conversational language.," *Am. J. Alzheimers. Dis. Other Demen.*, vol. 30, no. 1, pp. 61–8, Feb. 2015.
[237] E. Irazoki, J. A. García-Casal, J. Sánchez-Meca, and M. Franco-Martín, "Efficacy of group reminiscence therapy for people with dementia. Systematic literature review and meta-analysis," *Revista de Neurologia*, vol. 65, no. 10. Revista de Neurologia, pp. 447–456, 16-Nov-2017.
[238] N. A. Lazar, "Basic Statistical Analysis," in *The Statistical Analysis of Functional MRI Data*, 2008, pp. 1–36.
[239] A. E. Sarhan, "Estimation of the mean and standard deviation by order statistics," *Ann. Math. Stat.*, pp. 317–328, 1954.
[240] R. A. Groeneveld and G. Meeden, "Measuring Skewness and Kurtosis," *Stat.*, vol. 33, no. 4, p. 391, Dec. 1984.
[241] A. De Cheveigné, … H. K.-J. of the A. S. of, and undefined 2002, "YIN, a fundamental frequency estimator for speech and music," *asa.scitation.org*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.
[242] R. C. Sprinthall, *Basic Statistical Analysis*. .

[243] G. Corder and D. Foreman, “Nonparametric statistics: A step-by-step approach,” 2014.