applied sciences

MDPI

Article

Emotional Temperature for the Evaluation of Speech in Patients with Alzheimer's Disease through an Automatic Interviewer

Jesús B. Alonso-Hernández *, María Luisa Barragán-Pulido, Aitor Santana-Luis and Miguel Ángel Ferrer-Ballester

Instituto para el Desarrollo Tecnológico y la Innovación en Comunicaciones (IDeTIC), Universidad de Las Palmas de Gran Canaria, Despacho D-102, Pabellón B, Ed. de Electrónica y Comunicaciones, Campus de Tafira, 35017 Las Palmas, Spain; mbarragabl@idetic.eu (M.L.B.-P.); aitor.santana@ulpgc.es (A.S.-L.); miguelangel.ferrer@ulpgc.es (M.Á.F.-B.)

* Correspondence: jesus.alonso@ulpgc.es; Tel.: +34-928-452863

Abstract: In the context of the detection and evolutionary control of Alzheimer's disease from voice recordings and their automatic processing, this work aims to objectively determine the discriminatory capacity of a set of voice features linked to the emotional load of speech. We use descriptive statistics derived from the concept of emotional temperature as quantifiable characteristics of the voice. We apply a series of parametric and nonparametric analyses to the set of features, both individually and collectively, and explore their potential in relation to the use of different methods of unsupervised classification. With the aim of comparing how the type of interviewer used in the sample collection (i.e., voice recordings) influences the discrimination of AD through emotional speech analysis, we used the CSAP-19 database, which includes voice samples obtained through human interviewer (spontaneous speech samples) and automatic interviewer (induced speech samples) for the three defined populations (HC, mild AD, and moderate AD). In this regard, a comparative analysis is also conducted on the potential of emotional temperature features defined according to the sample collection process (manual or automatic interview process).

Keywords: Alzheimer's disease (AD); automatic interviewer; emotional temperature; telecare; telemedicine

Citation: Alonso-Hernández, J.B.; Barragán-Pulido, M.L.; Santana-Luis, A.; Ferrer-Ballester, M.Á. Emotional Temperature for the Evaluation of Speech in Patients with Alzheimer's Disease through an Automatic Interviewer. *Appl. Sci.* **2024**, *14*, 5588. https://doi.org/10.3390/app14135588

Academic Editors: Agnese Sbrollini and Aurora Saibene

Received: 13 May 2024
Revised: 16 June 2024
Accepted: 18 June 2024
Published: 27 June 2024

1. Introduction

The main cause of neurodegenerative dementia in the world today is Alzheimer's disease (AD), which accounts for 70–76% of dementia cases in developed countries with increasingly longer living populations [1]. Although the aetiology of AD is still unknown, it is known that its onset is insidious, it appears in adulthood, and it mainly leads to cognitive and behavioural impairments [2]. Memory loss is one of the first symptoms to appear and, gradually, other symptoms, such as difficulties with language use and temporal and spatial disorientations are added. In more advanced stages, the capability to perform daily activities or even basic body functions, such as walking or swallowing [3], decreases or disappears. In any case, when the first symptoms appear, the damage caused is irreparable and chronic. There is currently no cure for this disease, and it leads, for all intents and purposes, to neuronal impairment and death [4].

Today, the diagnosis process is unavoidably carried out in the advanced stages of the disease and is usually invasive and costly [5]. In this context, many studies have shown that speech analysis is a relevant indicator of a patient's emotional and cognitive states, and it can even detect the first symptoms years before a probable clinical diagnosis is established [6–8]. In recent years, techniques based on the automatic processing of the voice signal from a patient's record have found an important niche in language evaluation applied to the detection and monitoring of neurodegenerative diseases [9]. For this purpose, samples are classified according to different voice features using machine learning and

deep learning techniques [10]. It is worth noting that studies, in this regard, have generally based their feature extraction processes on the analysis of conventional parameters, namely, the duration of voiced and unvoiced segments, pitch, amplitude, and periodicity, as well as others obtained from frequency analyses and cepstral domains [11–13]. Nevertheless, despite these promising techniques, there are also certain limitations from the point of view of the linguistic differences that can be found, for example, in dialects of the same language or directly among different languages. From this perspective, several studies based on automatic speech analysis have progressed in the definition of new concepts linked to emotions, whose aim, among others, is to avoid this limitation [9].

Automatic emotion recognition from speech is of particular interest, since AD patients show changes in the way they express their emotions compared to cognitively healthy subjects. As the disorder develops, symptoms such as disorientation, mood changes, sleep disturbances, and confusion appear. Other symptoms that become progressively more noticeable include memory loss, behavioural changes, communication difficulties, and decreased motor skills [14].

Related Works

In recent years, several advanced methodologies and technologies, such as natural language processing, speech recognition, and machine learning, have been investigated for Alzheimer's disease (AD) detection through speech analysis. Different studies have been published analysing acoustic (prosody, voice quality, and pauses), lexical, and semantic features to identify biometric markers of AD using machine learning models, such as random forest [15], linear regression, deep neural networks [16], and ensemble methods, to classify AD status and predict scores on cognitive tests such as the MMSE. In addition, certain results have demonstrated significant associations with hippocampal volume and β-amyloid levels in cerebrospinal fluid, suggesting that these biomarkers may identify cognitive impairment in the preclinical and prodromal stages of AD and predict its progression [17]. These studies report high levels of accuracy in classifying AD and predicting cognitive impairment, with accuracies ranging from 80% to 90%, validating their models on public datasets, such as ADDReSS or AcceXible, as well as local datasets. Moreover, the approaches vary, ranging from the detection of pauses [18] in speech to the analysis of paralinguistic features and verbal fluency [19]. The results of the models are compared with benchmark algorithms, demonstrating significant improvements in accuracy. The research spans multiple languages and testing contexts, suggesting that speech analysis is a promising tool in the early detection of Alzheimer's disease, offering effective indicators of cognitive impairment and validation in diverse linguistic and cultural settings.

For its part, automatic emotion recognition from speech is of particular interest because AD patients exhibit changes in how they express their emotions compared to cognitively healthy subjects. Currently, there is no consensus on the number of emotions to be analysed [20]. Most research focuses on the following four basic emotions: anger, fear, sadness, and happiness [21]. In some studies, other emotions, such as surprise and disgust, have also been included [22]. Other research has focused on the development of real-time applications and emphasises the usefulness of representing emotions on an evaluation plane in terms of two or more continuous levels or dimensions [23]. In practice, the most commonly used dimensions to represent emotions are activation and valence levels [24]. The activation level relates to the perceived intensity of the emotion, and the valence level relates to the perceived pleasantness of a stimulus [24].

Numerous studies focusing on automatic emotion recognition from speech have carried out feature extraction based on prosodic aspects related to grammatical structure and lexical stress, such as phonation duration [25–27], pitch and energy contour, or Teager energy operator [28], related to grammar and lexical stress. Regarding paralinguistic aspects, some studies focus on other types of features, such as the first formant [26,27] or the energy concentration in different energy bands [29]. Localised studies about emotional speech analysis applied to AD are fundamentally based on three families of parameters:

acoustic features such as pitch (standard deviation (SD); maxima and minima) or intensity-related features (SD; maxima and minima), among others; voice-quality features, such as shimmer, local jitter, harmonic-to-noise ratio (HNR), noise-to-harmonic ratio (NHR), and autocorrelation, among others; and duration features such as voice and voiceless fragments. The short-term energy is the main feature analysed.

In recent years, a variety of research has explored the use of speech analysis and emotion recognition as noninvasive methods for the early detection of neurocognitive disorders such as Alzheimer's disease. These studies employ different datasets and methodologies, such as emotional prosody recognition (EPR), the ability to understand emotions through tone of voice [30], or the use of the Hurst exponent [31], to analyse speech signals in different languages, being able to effectively differentiate between emotions of anger and sadness, regardless of the language. Similarly, other studies assess emotions of frustration in picture description using speech emotion recognition (SER) to measure disease progression [32]. Intelligent and noninvasive computational techniques based on emotional feature extraction from speech and pattern recognition using neural networks have also been employed [33]. The results obtained in these studies are remarkable, showing a high sensitivity in distinguishing between healthy and cognitively impaired patients and among different degrees of disease severity. These studies suggest that deficits in emotional recognition are a part of AD symptoms and that speech analysis and emotion recognition offer promising noninvasive tools for the early diagnosis of Alzheimer's disease, potentially improving patients' quality of life through earlier and more accessible detection. These methods are effective regardless of the language, demonstrating their universal applicability and reduction in the costs and time compared to traditional methods.

In this context, on the basis of the temporal segmentation of the speech signal, the concept of emotional temperature (ET) has been defined [24]. This is a characteristic related to the emotional charge of speech calculated from various prosodic and paralinguistic aspects. Among the acoustic features of speech, the fundamental frequency of the samples is considered the main prosodic indicator, specifically the intonation given by the pitch contour. Furthermore, the accumulation of acoustic energy in different frequency bands, which varies according to the speech production model, can also be used as a paralinguistic indicator of emotional state. In emotional speech, the energy at higher frequencies increases compared to nonemotional speech. To calculate the ET, two prosodic features (related to pitch) and four paralinguistic features (related to energy) are extracted from each speech signal fragment [24].

From ET and the use of support vector machine (SVM) classifiers, some studies [24] have managed to classify speech samples as pathological and nonpathological, obtaining strong results for AD detection. In the same vein, combined emotional response analysis (ERA) methods have been developed that employ different linear features, such as pitch, intensity, or variation in the frequency components. Along with ET measures and spontaneous speech tasks, they have achieved discrimination between health control (HC) subjects and AD patients with high accuracy rates. Combining ASSA analysis [34,35] and features such as ET, it has been possible to demonstrate a significant loss of fluency in people with AD regarding the duration and percentage of voiced and unvoiced segments [35,36], with an accuracy of 92.24%. Combining ASSA, ESA, and ET, results can be achieved using SVMs with an accuracy of around 94% [34]. It is worth noting that ET has also been employed, together with the fractal dimension (FD), in some publications, in the context of the automatic analysis of emotional response (AAER) from spontaneous speech [37]. In that sense, they have demonstrated that the inclusion of fractal dimension features adds relevant information regarding the nonlinearity in speech signals and the appropriate analysis of emotional response [37]. In any case, studies based on ET, exclusively or in combination with other features, have shown promising results regarding the definition of useful features in the early diagnosis of AD.

Within the context of AD detection and evolutionary control from voice recordings and their automatic processing, this report aims to objectively determine the discrimina-

tive capacity of a set of voice features linked to the emotional speech charge from two different types of speech samples. First, we conceived this work as a preliminary approach to obtaining information on the potential of emotional features in the detection of AD. Second, another objective of this work was to conduct an exploratory study on the potential of speech samples collected through automated methods compared to their traditional counterparts (i.e., personal interviews). This last point is especially relevant considering the numerous advantages that the automatic collection of samples can offer a priori. Objectivity, scalability, speed, and low cost are some of these advantages. However, although there are some studies that have automated their interview process [12], the scarcity of systems that have applied it is one of the main reasons why it has not yet been possible to demonstrate to what extent the results obtained from an automated interview process are truly useful for the same subject [38]. To this day, it remains essential to understand how speech sample collection influences the detection of AD and its potential in the field of automatic voice analysis.

In order to carry out our study, we made use of descriptive statistics derived from the concept of emotional temperature, as a quantifiable voice feature. On this set of features, both individually and collectively, we applied a series of parametric and nonparametric statistics and studied their potential concerning the use of different unsupervised classification methods. In line with our previous studies, we started from the CSAP-19 database, which includes voice samples obtained by a human interviewer (spontaneous speech samples) and an automatic interviewer (induced speech samples) for the three defined populations (HC, mild AD, and moderate AD). In this sense, a comparative analysis was also carried out on the potentials of the defined emotional temperature features according to the sample collection process (manual or automatic interview process).

2. Materials and Methods

This section describes the methodology used to discriminate AD from the emotional charge of the voice of healthy subjects and AD patients. For this purpose, the Cross-Sectional Alzheimer Prognosis database released in 2019 (CSAP-19) [39] was used. This database is owned by University of Las Palmas de Gran Canaria and contains healthy and pathological samples obtained from the following two different types of interviewers that result in two different types of recordings: human and automatic interviewers. The first type of recording is characterised as having been made by an automatic interviewer using the Prognosis software v1, and the second type of recording, which is the most widespread in the field, was made by a research team member. In the following, we define the former as induced speech samples and the latter as spontaneous speech samples.

The automatic speech analysis was carried out using measures of emotional charge based on the emotional temperature of the subjects' vocalisations. Speech processing and feature extraction were carried out, obtaining a complete set of five measures for each sample. First, a univariate study was carried out using four statistics obtained from the ET variables under study and their subsequent parametric and nonparametric investigations based on the Wilcoxon test, Kruskal–Wallis test, and median test. To understand how the results varied using the complete set of features, we performed different multivariate analyses using classifications based on linear discriminant analysis (LDA), logistic regression, kth-nearest-neighbour (KNN) algorithms, and, finally, a multivariate parametric analysis (MANOVA). A process of individual feature selection was also carried out in order to understand the relevance of each feature within the complete set of features. Both univariate and multivariate statistics were applied to the three populations defined in the database: HC and mild and moderate AD.

2.1. Method

2.1.1. Calculation of Emotional Temperature

First, a voice activity detector (VAD) was applied to each of the recordings using MATLAB® software (R2019a). Once the different frames were obtained from the VAD, the

six following features were extracted: two prosodic features (related to pitch) and four paralinguistic features (related to spectral energy).

To calculate the ET, we used as prosodic parameters two linear regression coefficients (a and b) from Equation (1) that model the pitch contour $p(n)$ of an audio sample $\{w(n)\}$ [24].

$$Min(a,b) = \sum_{i=1}^{n} (p_i(n) - a - bx_i(n))^2 \quad (1)$$

Coefficients a and b were calculated using the least squares method. Coefficient a represents the pitch, while b is related to the tone trend. For this purpose, a pitch estimation algorithm (YIN) [40] was used.

To calculate the paralinguistic features, four voice spectral energy balances ($EB0$, $EB1$, $EB2$, and $EB3$) were used, which were quantified using four energy concentration percentages in four frequency bands, Bi (where $i \in [0, 3]$). Thus, with a sampling rate greater than 16 kHz, the frequency bands were divided into the following ranges: B_0 = [0 Hz–400 Hz], B_1 = [400 Hz–2 kHz], B_2 = [2 kHz–5 kHz], and B_3 = [5 kHz–8 kHz]. The percentage of energy in each EBi frequency band, in turn, was obtained using Equation (2).

$$E_{Bi} = \frac{\sum_{f=B_i} |X(f)|^2}{\sum_{f=0}^{8\text{kHz}} |X(f)|^2}, \text{where } 0 \leq i \leq 3 \quad (2)$$

where $|X(f)|^2$ corresponds to a period of the temporal voice frame, $w(n)$.

Once the six prosodic and paralinguistic features were obtained, to carry out the emotional temperature calculation, a support vector machine (SVM) classifier was used, specifically the LIBSVM library in MATLAB®. In the first phase, each emotional segment was classified into the following two types: high activation and low activation. The decision threshold (Th1) was calculated based on the equal error rate (EER) obtained from the training data. In the second phase, the complete speech signal was classified as high activation if the percentage of the emotional segments classified as high activation in the previous phase was higher than a second threshold (Th2). This second threshold (Th2) was calculated from the EER obtained from the validation data inserted into LIBSVM (which estimates the minimum percentage of high activation segments required for the signal to be classified as such). The resulting scale within this framework was linear and normalised. Generally, a voice signal was classified as high activation when ET $\geq$ 50 (emotional speech) and as low activation otherwise.

Finally, the discrete emotional temperature (ET_d), which is a number value, provides information based on the entire recording about whether the speech is emotional ($ET_d > 50$). On the other hand, the continuous emotional temperature segments the recording into different fragments and, from each one, obtains an emotional temperature value, thus obtaining a vector from which different descriptive statistics are calculated: mean value, variance, skewness, and kurtosis.

2.1.2. Descriptive Statistics

For each sequence based on speech, the measures of ET_d and their correspondent descriptive statistics measures (i.e., mean value [41], variance [42], skewness, and kurtosis [43]) were extracted, obtaining, specifically, the average ET $\overline{(et_s)}$, variance of the ET $\left(\sigma^2_{et_S}\right)$, skewness of the ET $\left(\widetilde{\mu}_{et_{S_3}}\right)$, and kurtosis of the ET $(Kurt_{et_S})$.

- ET_d: the discrete emotional temperature of the recording (see Section 2.1.1);
- Average of the continuous emotional temperature $\overline{(et_s)}$: refers to the continuous emotional temperature vector values and describes the mean value of the different ET values of the sound fragments in a recording. It is estimated using the following estimator of the arithmetic mean [42]:

$$\overline{(et_s)} = \frac{\sum_{i=1}^{N} et_{c_i}}{N} \quad (3)$$

where et_{c_i} is the value of et_c for each fragment $(S_1, S_2, \ldots, S_N)$ into which each voice recording $\{S_i\}$ is divided;

- Variance in the continuous emotional temperature $\left(\sigma^2_{et_c}\right)$: refers to the continuous emotional temperature vector values and describes the variation in the different fragments in a recording. It is estimated using the following estimator of the variance [42]:

$$\sigma^2_{et_c} = \frac{\sum_{i=1}^{N} \left(et_{c_i} - \overline{et_c}\right)^2}{N-1} \tag{4}$$

- Skewness of the continuous emotional temperature $\left(\widetilde{\mu}_{et_{S_3}}\right)$: refers to the continuous emotional temperature vector values. This measure allows for characterising the behaviour of the probability distribution function of the ET values of the different fragments. This measure quantifies [43] the lack of symmetry of the average ET values of the voice fragments. Positive or negative values of $\widetilde{\mu}_{et_{c3}}$ indicate data skewed to the right of their distribution curve or to the left, respectively. The skewness of ET of speech is calculated using the following estimator:

$$\widetilde{\mu}_{et_{S_3}} = \frac{\sum_{i=1}^{N} \left(et_{c_i} - \overline{et_c}\right)^3}{N \cdot \left(\sqrt{\sigma^2_{et_c}}\right)^3} \tag{5}$$

where et_{c_i} is the ET value of each sound fragment, $\overline{et_c}$ is the average of the ET values, $\sigma^2_{et_c}$ is the variance in the ET values, and N is the number of sound fragments in the sample of speech;

- Kurtosis of the continuous emotional temperature ($Kurt_{et_c}$): refers to the continuous emotional temperature vector values. This is a measure that allows for characterising another aspect of the behaviour of the probability distribution function of the ET values of the different fragments. This measure states the quantity of sound fragments in a recording with an ET value that is close to the average ET $\left(\overline{et_c}\right)$. The larger the value of $Kurt_{et_c}$, the steeper its distribution curve. $Kurt_{et_c}$ is calculated using the following estimator [43]:

$$Kurt_{etc_s} = \frac{\sum_{i=1}^{N} \left(etc_{S_i} - \overline{etc_S}\right)^4}{N \cdot \left(\sqrt{\sigma^2_{etc_S}}\right)^4} \tag{6}$$

Once the results of the emotion study were obtained for each sample, the five variables were stored in a single text document. From these measures, univariate and multivariate analyses were performed using Stata® software, version 13.0 [44].

2.1.3. Univariate Analysis

To carry out the analysis, the samples were classified according to each of the three populations under study (HC, AD1, and AD2). The population referred to as AD included samples of mild and moderate grades (AD1 and AD2 groups).

Since we did not know a priori whether the samples under study followed a normal distribution, we started by performing a parametric analysis using the linear regression method and its subsequent residual analysis to check the suitability of the parametric analysis for our study. The normality tests of the residuals performed were the skewness and kurtosis tests, where, from the Chi-square values, we can determine if the regression is correct [45].

Subsequently, on the basis of the results obtained, a descriptive statistical analysis was carried out (specifically on the basis of the mean and standard deviation values of the ET variables) and a nonparametric study based on the Wilcoxon rank sum test [46], Kruskal–Wallis test [47], and median test. For any of these tests, $Prob|z|$ values greater than

0.05 ($Prob\,|z| > 0.05$) were cases in which there was no difference found when comparing samples from different populations.

2.1.4. Multivariate Analysis

To analyse the complete set of emotional features as a single set and not as individual features, the classification methods used were linear discriminant analysis (LDA), logistic regression, and kth-nearest neighbour (KNN) discriminant analysis.

These classifiers were chosen because of the diversity of approaches they offer. Each of these algorithms has a different approach to classification. For instance, LDA aims to maximise the separation among classes based on the features, while the logistic regression classifier models the probability of belonging to a class. On the other hand, the k-NN classifier is based on the similarity with the nearest neighbours.

Each of the classifiers were applied to the five defined emotional measures (et_d, $\overline{et_S}$, $\left(\sigma^2_{et_S}\right)$, $\widetilde{\mu}_{et_{S3}}$, and $Kurt_{et_s}$) following the leave-one-out cross-validation technique (LOOCV), with the aim of maximising the amount of data used to train the model in each iteration. Specifically, for the KNN classifier, classification was carried out for the three different scenarios by choosing 1, 3, and 5 nearest neighbour samples.

In addition to the previous classifiers, a multivariate analysis of variance (MANOVA) was performed. In doing so, it was previously checked whether the samples followed a Gaussian distribution and whether it was appropriate to conduct this type of analysis. The hypothesis test used was based on the measures of skewness.

For all of the mentioned classifiers, two types of classifications were performed according to the different populations. Thus, the first type of classification was based on the presence or absence of AD and the second type was based on the different degrees of the disease defined in this study (mild AD, moderate AD, and HC). In the case of classifications based on the presence or absence of the disease, the values were obtained by comparing HC subjects with mild and moderate AD subjects. Finally, for all scenarios and classifiers, confusion matrices were calculated together with the values for sensitivity, specificity, and accuracy.

2.1.5. Feature Selection

To determine the relevance of each of the emotion variables analysed in this study, a feature selection process was carried out using neighbourhood component analysis for classification. Specifically, the function fscnca, developed for MATLAB, was applied to the five analysed emotion features. This function assigns a weight to each feature within the set using a diagonal adaptation of neighbourhood component analysis (NCA).

Additionally, the impact of the type of interviewer employed was examined, as well as the influence of the number of classes used, as follows: disease or degrees.

2.2. Materials

Database

To carry out this study, recordings contained in the Cross-Sectional Alzheimer Prognosis R2019 database [39] were used. This database was created to assess how discriminant a voice sample can be according to the type of interviewer used in a recording. It consists of the following two types of recordings, according to the type of interviewer with whom the samples were obtained, as follows: an interview process in which the subject is invited to speak freely for a few minutes (i.e., spontaneous speech) or using Prognosis software [39], in what has been called induced speech.

In general, each subject provided a total of 4 voice recordings, with three induced speech recordings and one spontaneous speech recording. The average duration of the recordings was 34.5 s, with a sampling frequency of 44,100 Hz and in a WAV file format. For the recordings, a computer with detailed specifications (Intel Core i7, 6 GB RAM, 750 GB hard drive, and a 16.9-inch LCD monitor) was used along with Ozone Rage ST headphones with a microphone, providing stereo sound, 32 Ω impedance, frequency range from 20 to

20 kHz, and a microphone with specific characteristics of impedance, sensitivity, directivity, and frequency response.

The recordings of spontaneous speech were conducted by a member of the research team using a laptop and stored digitally. The Audacity® software, version 3.0, was used along with Ozone Rage ST headphones with a microphone. During the interviews, subjects were encouraged to speak freely on any topic to obtain spontaneous speech recordings, ranging in duration from 30 s to 2 min. The samples obtained with the automated interviewer were conducted using the Prognosis software [39], the same laptop, and Ozone Rage ST headphones with a microphone, as used with the human interviewer. Three induced speech recordings were obtained per participant.

The database contains a total of 87 recorded subjects, as follows. There were 41 AD subjects and 46 HC subjects, all over 65 years of age. Sixty-four percent of the participants were women compared to thirty-six men. According to the grade of the disease, among the 41 patients with AD, 26 corresponded to the mild grade and 15 to the moderate grade [39].

3. Results

3.1. Univariate Analysis

3.1.1. Descriptive Statistical Analysis

This section presents the descriptive statistical results expressed in terms of the mean and standard deviation for the five emotional temperature variables previously defined. A categorisation of the samples was carried out considering each of the three populations under investigation (i.e., HC, AD1, and AD2). The population designated as AD included the samples corresponding to the mild and moderate grades. Table 1 presents the values of the variables obtained by both interviewers.

Table 1. Descriptive statistical values of the emotional temperature measures for each population and for each interviewer: mean value (μ) and standard deviation (σ).

	Populations							
	HC *		**AD1 ***		**AD2 ***		**AD (AD1 + AD2)**	
Variable/ Interview	**Human $\mu(\sigma)$**	**Automatic $\mu(\sigma)$**	**Human $\mu(\sigma)$**	**Automatic $\mu(\sigma)$**	**Human $\mu(\sigma)$**	**Automatic $\mu(\sigma)$**	**Human $\mu(\sigma)$**	**Automatic $\mu(\sigma)$**
et_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)
$\overline{et_c}$	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)
$\sigma^2_{et_c}$	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)
$\widetilde{\mu}_{et_{c3}}$	−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)
$Kurt_{et_c}$	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: healthy control; AD1: mild-grade Alzheimer's; AD2: moderate-grade Alzheimer's.

At first glance, similarities can be seen in the values obtained by the human interviewer and the automatic interviewer for some of the variables analysed, such as for the variable et_d.

3.1.2. Parametric Analysis

The results obtained do not correspond to a normal distribution, which rules out the suitability of carrying out a parametric analysis in this case. Examining the linear regressions of all emotional temperature variables, it is evident that the regression line does not adequately fit the totality of the data. The residual normality tests, specifically the skewness and kurtosis tests, indicate, according the chi-square values, that the regression is not correct and the residuals would not follow a normal distribution.

3.1.3. Nonparametric Analysis

Table 2 presents the results obtained from the three nonparametric tests performed to compare the human interviewer with the automatic interviewer. Values for which $Prob\,|\,z\,|$ is higher than 0.05 ($Prob\,|\,z\,| > 0.05$) are highlighted in grey, indicating the absence of differences when comparing the populations' measures. These results support the null hypothesis, suggesting that a variable is not discriminant for Alzheimer's disease (AD). It is observed that, for this type of emotion variable, a discrimination among populations is not so evident. However, residually, we found some values that could lead to the rejection of the null hypothesis, as follows: et_d and $\widetilde{\mu}_{et_{c3}}$ for the automatic interviewer and $Kurt_{et_c}$ for the human interviewer.

Table 2. Results of the univariate nonparametric analysis: discriminant capacity of the different populations and interviewers in relation to the emotional temperature variables.

	Wilcoxon Test		**Kruskal–Wallis Test**		**Median Test**	
	Prob | z |		**χ^2**		**Pearson χ^2**	
Variable/ Interviewer	**Human**	**Automatic**	**Human**	**Automatic**	**Human**	**Automatic**
HC * vs. AD						
et_d	0.34	0.05	0.342	0.05	0.56	0.12
$\overline{et_c}$	0.61	0.19	0.634	0.21	0.68	0.18
$\sigma^2_{et_c}$	0.81	0.08	0.824	0.09	0.93	0.28
$\widetilde{\mu}_{et_{c3}}$	0.79	0.06	0.791	0.06	0.93	0.05
$Kurt_{et_c}$	0.13	0.55	0.129	0.55	0.37	0.96
HC vs. AD1 *						
et_d	0.17	0.05	0.17	0.05	0.26	0.13
$\overline{et_c}$	0.47	0.28	0.50	0.30	0.66	0.23
$\sigma^2_{et_c}$	0.74	0.25	0.76	0.28	0.66	0.36
$\widetilde{\mu}_{et_{c3}}$	0.60	0.02	0.60	0.02	0.93	0.03
$Kurt_{et_c}$	0.34	0.93	0.34	0.93	0.66	0.36
HC vs. AD2 *						
et_d	0.41	0.27	0.41	0.27	0.60	0.46
$\overline{et_c}$	0.69	0.30	0.71	0.32	0.60	0.57
$\sigma^2_{et_c}$	0.12	0.06	0.14	0.07	0.60	0.35
$\widetilde{\mu}_{et_{c3}}$	0.54	0.85	0.54	0.s8483	0.60	0.85
$Kurt_{et_c}$	0.05	0.16	0.05	0.16	0.12	0.35
AD1 vs. AD2						
et_d	0.08	0.75	0.08	0.75	0.12	0.59
$\overline{et_c}$	0.52	0.87	0.55	0.87	0.53	0.89
$\sigma^2_{et_c}$	0.11	0.32	0.14	0.34	0.53	0.79
$\widetilde{\mu}_{et_{c3}}$	0.18	0.20	0.18	0.20	0.65	0.28
$Kurt_{et_c}$	0.24	0.14	0.24	0.14	0.12	0.08

* HC: healthy control; AD1: mild-grade Alzheimer's; AD2: moderate-grade Alzheimer's. Values where Prob | z | exceeds 0.05 (Prob | z | > 0.05) are shaded in grey, indicating no significant differences when comparing the populations' measures.

3.2. Multivariate Analysis

Using the five established emotional temperature variables and applying LDA classifiers, logistic regression, KNN (for n = 1, n = 3, and n = 5), and MANOVA to the database, two types of classifications were carried out, as follows: consideration of the presence or absence of AD and consideration of the different degrees of the disease (mild and moderate AD).

3.2.1. Multivariate Classification Based on the Presence or Absence of Disease

Table 3 presents the results, represented in a confusion matrix, for each interviewer and classifier. This table provides information about the number of classified samples and their percentage of the total.

Table 3. Confusion matrix of the multivariate classification using LDA, logistic classifier, and KNN based on the presence or absence of disease.

		Automatic Interviewer			Human Interviewer		
Classifier	**True Disease**	**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%)
Logistic	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%)

Grey color—sum of values.

The results of the confusion matrix in Table 3 are translated into the accuracy, sensitivity, and specificity for each interviewer and presented in Table 4. Although the performance was lower compared to previous studies conducted on the time variables [48], the best results are observed for the logistic classifiers, both for the human and the automatic interviewers.

Table 4. Accuracy, sensitivity, and specificity values for the human and automatic interviewers based on the multivariate classification using LDA, logistic classifier, and kNN according to the absence or presence of disease.

	Classifier	**Accuracy [%]**	**Sensitivity [%]**	**Specificity [%]**
Automatic interviewer	LDA	55.82%	57.66%	54.35%
	Logistic	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%
Human interviewer	LDA	47.89%	48.00%	47.83%
	Logistic	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%

3.2.2. Multivariate Classification Based on Different Grades of the Disease

Table 5 presents the results, represented in a confusion matrix, for each interviewer and classifier, according to the following different grades defined for the disease: no disease (0), mild (1), and moderate (2). This table provides information on the number of classified samples and their percentage of the total.

Table 5. Confusion matrix of the multivariate classification results using LDA, logistic classifier, and KNN based on different grades of disease: no disease (0), mild (1), and moderate (2).

Classifier	True Grade	Automatic Interviewer				Human Interviewer			
		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%)
Logistic	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%)

Grey color—sum of values.

Table 6 translates the values from the confusion matrix into values of accuracy, sensitivity, and specificity obtained for each of the classifiers and interviewers. From them, it can be extracted that, once again, the best results were achieved with the logistic classifiers independently of the interviewer used.

Table 6. Accuracy, sensitivity, and specificity values for the automatic and human interviewers based on the multivariate classification using LDA, logistic classifier, and kNN for different disease grades: no disease (0), mild disease (1), and moderate disease (2).

	Classifier	Accuracy [%]	Sensitivity [%]	Specificity [%]
Automatic interviewer	LDA	38.55%	63.96%	46.38%
	Logistic	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%
Human interviewer	LDA	40.85%	56.00%	39.13%
	Logistic	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%

3.2.3. Multivariate Classification MANOVA

First, it was fundamental to verify whether the samples followed a Gaussian distribution. For this purpose, we performed a normality test using measures of skewness and kurtosis.

Since it was confirmed that the results of this assessment supported the application of a MANOVA analysis, the *p*-values obtained from the multivariate analysis are presented in Table 7, in which W, P, L, and R correspond to Wilks' lambda, Pillai's trace, Lawley–Hotelling trace, and Roy's largest root statistics, respectively. In grey, the statistics that suggest that the set of variables does not distinguish among the different AD groups are highlighted. All four statistics indicate that the dependent variables analysed discriminate only between the HC and AD populations. Consequently, when the grouping variable is based on the grade of disease, the set of emotional temperature variables is no longer discriminatory for AD.

Table 7. Multivariate MANOVA analysis: comparison between automatic and human interviewers and the different AD populations. Wilks' lambda, Lawley–Hotelling trace, Pillai's trace, and Roy's largest root statistics applied to the set of emotional temperature measures.

	MANOVA							
	Disease (HC *-AD)		Grade (HC-AD1 *)		Grade (HC-AD2 *)		Grade (AD1-AD2)	
	p-Value		*p*-Value		*p*-Value		*p*-Value	
Statistic/ Interviewer	H	A	H	A	H	A	H	A
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: healthy control; AD1: mild-grade Alzheimer's; AD2: moderate-grade Alzheimer's; W: Wilks' lambda; P: Pillai's trace; L: Lawley–Hotelling trace; R: Roy's largest root; H: human interviewer; A: automatic interviewer. The statistics indicating that the set of variables does not differentiate among the different AD groups are highlighted in grey.

3.3. Feature Selection

As can be seen in Table 8, the results obtained during the feature selection process remained relatively constant, with the kurtosis and skewness features being irrelevant regardless of the interviewer used and the type of classification.

Table 8. Relevance analysis of emotional temperature features (A: automatic interviewer; H: human interviewer).

Classification	Interviewer	te_d	$\overline{te_c}$	$\sigma^2_{te_c}$	$\tilde{\mu}_{te_{c3}}$	$Kurt_{te_c}$
Based on absence or presence of disease	Automatic	0.9	2.1	0.15	0	0
	Human	0.55	1	0.35	0	0
Based on different grades of disease	Automatic	0.4	1.7	0.1	0	0
	Human	0.45	0.95	0.3	0	0

In this way, we decided to use the set of samples from the automatic interviewer and the binary classification as a reference for our analysis. Table 9 shows the results of the emotional temperature feature selection process under these conditions.

Table 9. Feature selection: results of the emotional temperature measure selection.

Emotional Feature	Relevance
te_d	0.9
$\overline{te_c}$	2.1
$\sigma^2_{te_c}$	0.15
$\tilde{\mu}_{te_{c3}}$	0
$Kurt_{te_c}$	0

Based on the values presented in Table 9, the results of the feature selection process applied to emotional temperature measures are displayed in Figure 1, visually representing the weight of each feature in the set.

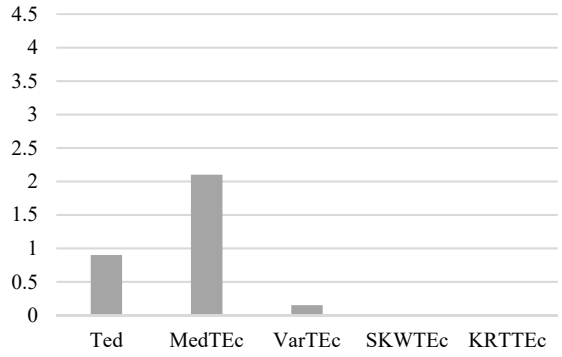

Figure 1. Graphical results of the feature selection process applied to the emotional temperature measures.

The results indicate that the best measures, by a wide margin over the other features, were TEd (te_d) and MediaTEc ($\overline{te_c}$). Although, to a lesser extent, VarTEc ($\sigma^2_{te_c}$) also proved to be relevant to the set. In any case, the most relevant feature was MediaTEc ($\overline{te_c}$).

4. Discussion

The study of speech in patients with Alzheimer's disease using automatic speech processing is carried out primarily through face-to-face interviews, in which an interviewer asks questions, assigns speech tasks, or stimulates the subject to speak. Databases containing samples obtained from fully automated interview processes represent a small part of the total available. Despite the many advantages identified in automated sample collection, the number of systems currently implementing it remains limited. At present, this limitation could be one of the main barriers to demonstrating the usefulness of the results obtained by these automated methods. In this sense, exploring the potential of such tools and how they can be progressively improved are essential tasks in the field of automatic speech analysis for the detection of Alzheimer's disease.

To gain a better understanding of the possible contribution of automatic interviewers applied to the discrimination of Alzheimer's disease through speech, several recordings obtained from both human and automatic interviewers were compared. Using the CSAP-19 database, as presented above, and considering the previously demonstrated usefulness of speech timing analysis, an emotional temperature feature extraction process was performed on the two types of speech present in the database. In this study, several univariate and multivariate analyses were carried out on the three populations included in the CSAP-19, as follows: healthy control subjects (HC), mild Alzheimer's (AD1), and moderate Alzheimer's (AD2).

Following previous publications on several temporal measures of speech, we first analysed the five defined variables individually, as follows: discrete emotional temperature (et_d), average of continuous emotional temperature ($\overline{et_c}$), variance of continuous emotional temperature ($\sigma^2_{et_c}$), skewness of continuous emotional temperature ($\widetilde{\mu}_{et_{c3}}$), and kurtosis of continuous emotional temperature ($Kurt_{et_c}$).

From the results shown in Table 1, in Figure 2 the values obtained for each interviewer, grade, and defined variable are represented (human interviewer: dark grey; automatic interviewer: light grey). From the first analysis carried out, based on the analysis of the descriptive statistics, similar values and trends were obtained for the variables ETd (et_d), SKWETc $\left(\widetilde{\mu}_{et_{c3}}\right)$, and KRTETc ($Kurt_{et_c}$) independently of the interviewer used. On the other hand, for the five characteristics analysed, the HC subjects shared the same starting point regardless of the interviewer. Specifically, the variable ETd (et_d) for the human interviewer showed a drop in the mean values of emotional temperature in the AD2 patients compared to HC subjects. The KRTETc ($Kurt_{et_c}$) values less than three reflect less data concentrated around the average ET value. For AD2 patients, this fact is reflected to a greater degree for both interviewers. In general, regarding the rest of the variables, no clear trend or pattern was identified. In that sense, to detect it, another more exhaustive type of analysis is necessary.

From the different nonparametric analyses carried out (Wilcoxon test, Kruskal–Wallis test, and median test), we were able to determine which of the five variables considered would be discriminant for Alzheimer's disease using four different comparisons among the populations. According to the results in Table 2, the variables that showed the best results were ETd and SKWETc (interviewer: automatic; populations: HC/AD). In the case of the ETd variable, it was close to the limits established as the maximum ($Prob|z| \leq 0.05$). Likewise, the null hypothesis was also rejected for the SKWETc variable if we compare the HC/AD1 populations. Regarding the human interviewer, only the KRTETc variable would reject the hypothesis with a p-value equal to 0.05 in the Wilcoxon and Kruskal–Wallis tests. This indicates that there is a 5% probability of obtaining some difference between the HC and AD2 groups if we consider all groups to be equal.

The AverageETc and VarETc variables were not discriminant in any of the four scenarios studied. This does not suggest that these variables are not relevant for discrimination but, more specifically, there is not enough evidence to reject the null hypothesis. Concerning the comparison between the AD1 and AD2 populations, no result yielded a p-value higher than 0.05. This may be because the samples for the AD1 and AD2 populations were relatively small in number and similar to each other. Specifically, the AD2 population has a smaller number of samples in the database, with a total of 15 patients registered as AD2 patients. From the AD1 population, there are recordings of 26 patients and a total of 46 healthy control subjects interviewed. Therefore, it may be recommended to increase the number of samples, especially in the AD2 group, to obtain more consistent and conclusive results in this regard.

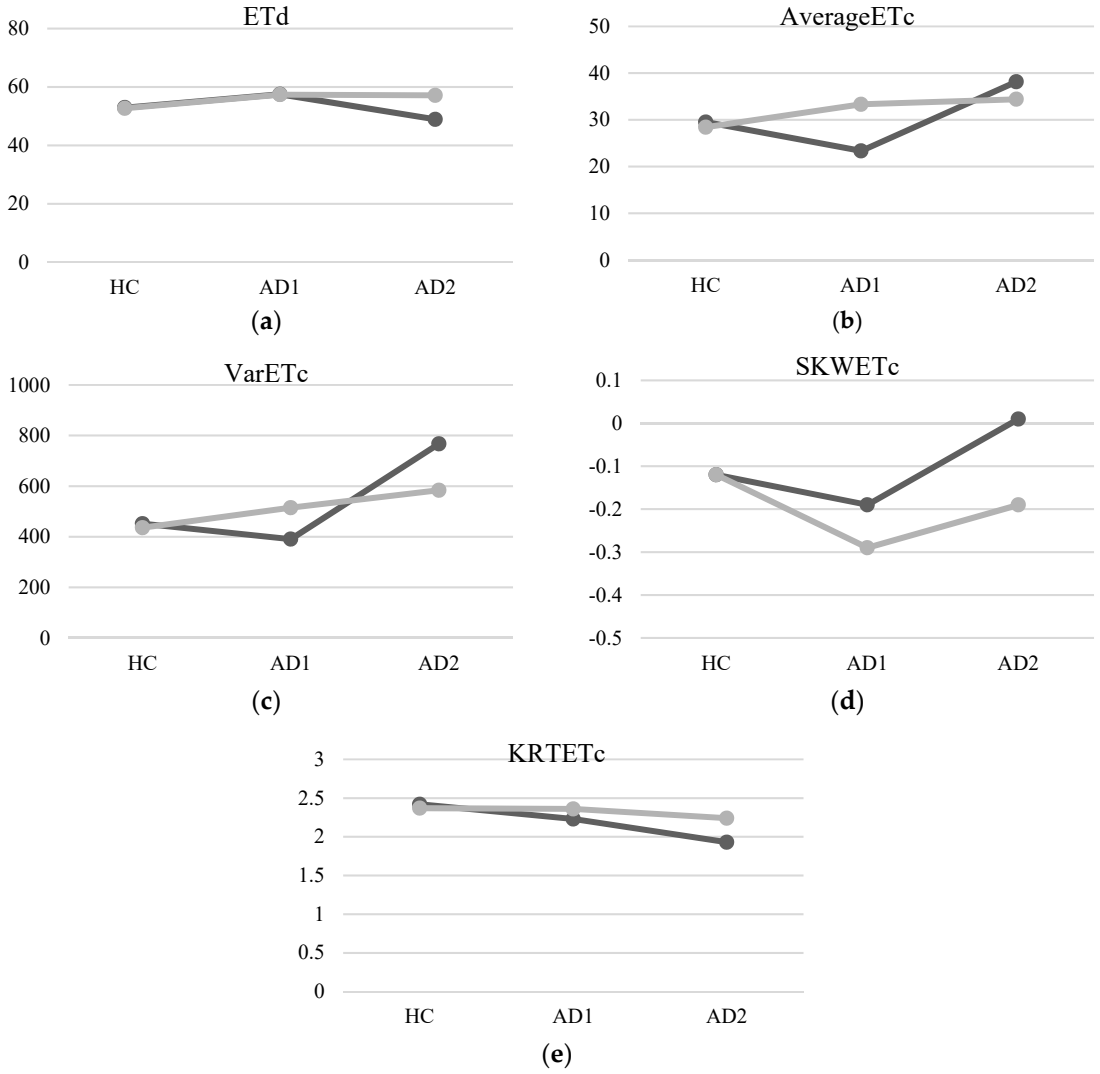

Figure 2. (**a**) Comparison between ETd $(\mathrm{et_d})$ values for human and automatic interviewers; (**b**) comparison between the average ETc $(\overline{\mathrm{et_c}})$ values for human and automatic interviewers; (**c**) comparison between VarETc $(\sigma^2_{\mathrm{et_c}})$ values for human and automatic interviewers; (**d**) comparison between SKWETc $(\widetilde{\mu}_{\mathrm{et_{c3}}})$ values for human and automatic interviewers; (**e**) comparison between KRTETc $(\mathrm{Kurt_{et_c}})$ values for human and automatic interviewers.

In turn, we have carried out a multivariate analysis using three different classifiers, as follows: LDA, logistic classifier, and KNN classifier (for values $n = 1$, $n = 3$, and $n = 5$). In Figure 3a,b,e,f, obtained in Tables 4 and 6, it can be seen that, when considering these three classifiers, the most favourable results in terms of accuracy and specificity were obtained for the binary classification (healthy–pathological). In both the automatic interviewer mode (59.4%) and the human interviewer mode (57.8%), the accuracy peaked using the logistic classifier. In terms of the specificity values, the highest values were achieved using the KNN classifier, for n = 3 and n = 5, with the human interviewer (69.6%) and the automatic interviewer (62.3%), respectively. In this sense, three different performance metrics were used for model evaluation: accuracy, sensitivity, and specificity. Accuracy is the proportion of correct predictions out of the total instances evaluated, sensitivity is the proportion of true positives out of the total instances that are actually positive, and specificity is the proportion of true negatives out of the total instances that are actually negative.

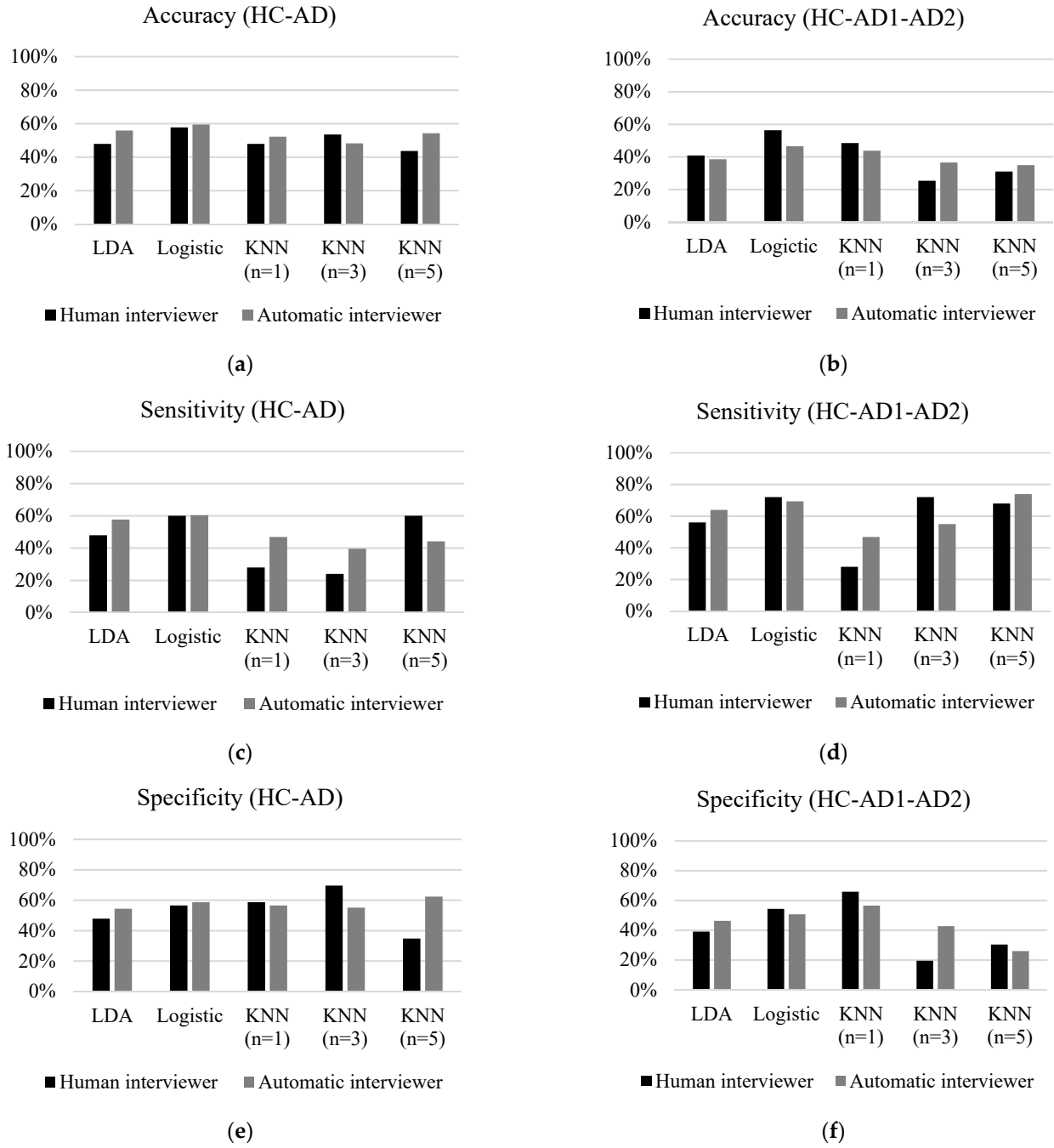

Figure 3. (**a**) Accuracy for the different classifiers: human and automatic interviewers; classification by AD presence or absence. (**b**) Accuracy for the different classifiers: human and automatic interviewers; classification by AD grades. (**c**) Sensitivity of the different classifiers: human and automatic interviewers; classification by AD presence or absence. (**d**) Sensitivity of the different classifiers: human and automatic interviewers; classification by AD grades. (**e**) Specificity for the different classifiers: human and automatic interviewers; classification by AD presence or absence. (**f**) Specificity for the different classifiers: human and automatic interviewers; classification by AD grades.

Comparing the two classification methods (by disease and by grade), there was a notable difference in the sensitivity values (Figure 3c,d) as opposed to the accuracy and specificity. The results show that classification by grade outperformed classification by absence or presence of disease, since their sensitivity values were considerably higher. In particular, the best values were above 70% for both types of interviewers: human (with a

sensitivity of 72% for the logistic and KNN classifier, n = 3) and automatic (with a sensitivity of 73.9% for the KNN classifier with n = 5). Nevertheless, it is crucial to point out that the sensitivity values obtained, regardless of the classifier used, were low. Specifically, for the case where sensitivity peaks with the automatic interviewer, we found that specificity was remarkably low (26.1%).

Although our main interest lies in sensitivity, which is the ability to detect subjects with pathology, a low specificity indicates an unacceptable excess of false positives in the classification. The accuracy reflects the dispersion of sample values and refers to the ratio of correct predictions to total predictions. Conceptually, we must consider the "cost" associated with each type of classification error in the algorithm when deciding which parameter to prioritise. Specifically, for our study, the detection of pathological subjects is crucial, so sensitivity should be the main parameter for evaluating our classifiers. However, as discussed, it is also necessary to find a balance between sensitivity and specificity to achieve a truly balanced system.

Overall, the best-performing classifiers were KNN (for n = 3 and n = 5) and the logistic classifier. In terms of sensitivity values, both interviewers achieved their best results when classifying by grade, with values around 70%. When analysing the results, from the point of view of the interviewer, neither of the two interviewers stands out from the other in any of the parameters evaluated.

The last multivariate analysis carried out in this study was the MANOVA variance test. The data obtained and presented in Table 7 indicate that, in all statistical aspects evaluated, the set of emotional temperature features is discriminant with the HC and AD populations and in the case of employing the automatic interviewer.

However, when considering the grouping variables AD1 and AD2, the statistic values for both interviewers are well above the upper limit set for the p-value ($Prob\,|z| = 0.05$). This phenomenon might be in line with the results of the nonparametric tests performed earlier. In these tests, the HC-AD discrimination showed a higher number of variables that met the limit established for the p-value. Specifically, in the three univariate nonparametric analyses carried out, none of the five variables independently proved to be discriminatory when comparing the AD1 and AD2 populations. This situation can again be attributed to the number of samples available for each population (the more samples, the more conclusive the results), as well as the additional difficulty in differentiating between two pathological voices of different grades. As can be deduced from the results obtained, the samples that would reject the null hypothesis in terms of AD discrimination are those generated by induced speech, i.e., those obtained through the automatic interviewer.

In general terms, in this study, we have evaluated the discrimination capacity between a set of samples obtained using an automatic interviewer and another set obtained using a human interviewer. All samples have been classified based on the emotional temperature measures extracted from them.

As can be inferred from the data presented, the samples show a similar pattern in their measures regardless of the interviewer used, being able to discriminate AD from induced or spontaneous speech. That can be seen both in the univariate analysis of the descriptive statistics (SKWETc and KRTETc) and in the behaviour and results of the different classifiers used (especially in terms of accuracy, but also in sensitivity and specificity).

These results are relevant because, once extracted and analysed, we can suggest that there is room for improvement in several aspects, such as the analysis methods used, the parameterisation process and even the tool used for the collection of induced speech samples, the Prognosis software (for example, to obtain more natural recordings from the subjects). Concerning the parameterisation process, in addition to the emotional temperature variables analysed, it would be particularly interesting to be able to analyse them in conjunction with the temporal measures already presented in previous works to gain a deeper understanding of the behaviour of the automatic interviewer and its potential.

This work constitutes a first approach to defining and studying the benefits of automatic sample collection compared to manual methods. It presents a series of objective

data for comparing both types of samples based on five emotional speech characteristics. Based on the results presented in this study, the capacity of these automatic techniques can be preliminarily and objectively assessed, which, though still under study, seem to have untapped potential. Expanding and deepening studies of this type is of particular importance given that the diagnosis of AD remains an unresolved issue, with up to 90% of mild cases potentially undiagnosed. Current diagnostic methods are costly and invasive for patients. Automatic solutions for the early diagnosis of AD using voice analysis would undoubtedly help democratise its administration. It would be interesting to develop web applications based on these linguistic biomarkers, potentially being used as a screening method for a large population, in addition to other benefits such as reducing the burden on healthcare personnel and systems.

5. Conclusions

Currently, the automation of the interview process applied to the discrimination of healthy subjects from those suffering from Alzheimer's disease through automatic voice processing is still an unexplored and promising field. This type of method has many significant benefits, such as its capacity for early detection and its evident advantages as a screening method; it is easy to apply, noninvasive, inexpensive and does not require the assistance of medical specialists. However, pending more conclusive results, neither speech analysis for the detection of Alzheimer's disease nor an automatic speech sample recording process are widely used in the probable diagnosis.

In this context, it is of particular interest to assess whether the process of automating AD interviews can provide sufficiently accurate samples to distinguish between healthy and ill subjects. If so, the advantages of automated interviews over their manual counterparts would extend from the objectivity and replicability of the process, the ease of monitoring and scalability of each patient's treatment, to the comfort for the interviewee and the possibility, where appropriate, of anonymity.

To this end, we carried out a detailed study focused on the extraction and parameterisation of emotional temperature features from a set of samples obtained using a human interviewer and an automatic interviewer. For this purpose, we carried out a series of univariate and multivariate analyses to determine the discriminative capacity of each type of interviewer and sample (spontaneous speech and induced speech). From the different analyses, the results obtained by the automatic interviewer, both in terms of accuracy and sensitivity and specificity (similar in many cases to those obtained from the human interviewer), reveal a mutual pattern regardless of the interviewer used, not being able to discern which one allowed a more accurate classification. In this sense, with a view to future works, it is suggested to explore combinations with other types of features, not only emotional ones. These parameters could be, for example, temporal (which have already demonstrated their relevance when measuring cognitive impairment and, specifically, in the detection of AD), frequency or cepstral to go deeper into the behaviour and capacity of the automatic interviewers. On the other hand, another future line would be to apply different classification methods to the obtained samples and different data models for a comparative analysis. Of course, continuing to improve the tool used for sample collection, the Prognosis software, is a task that would be of interest to work on in order to make interactions with participants more natural, for example, by developing processes of active listening. Finally, it would be interesting to develop web applications based on these linguistic biomarkers for evolutionary and pharmacological control through voice signal to facilitate the screening and monitoring process from home. Likewise, another future line would involve further expanding the database on which our study is based, taking new samples both cross-sectionally and longitudinally over time.

In any case, it seems clear that the efficient automation of interview processes to detect and monitor the progression of Alzheimer's disease represents a step forward in the current development of speech-based eHealth 4.0 solutions. This relatively new solution democratises the evolutionary monitoring of the disease and offers a faster, more accessible

and scalable alternative (telecare). They also provide additional objective parameters, which can be used to complement other methods currently in use, simplifying the work of the specialist physician and without requiring significant additional infrastructure.

Author Contributions: Conceptualisation, J.B.A.-H. and M.L.B.-P.; methodology, J.B.A.-H.; software, J.B.A.-H.; validation, J.B.A.-H., A.S.-L. and M.L.B.-P.; formal analysis, M.L.B.-P. and A.S.-L.; investigation, A.S.-L. and M.L.B.-P.; resources, J.B.A.-H.; data curation A.S.-L. and M.L.B.-P.; writing—original draft preparation, M.L.B.-P.; writing—review and editing, A.S.-L. and M.L.B.-P.; visualisation, M.L.B.-P.; supervision, J.B.A.-H. and M.Á.F.-B.; project administration, J.B.A.-H. and M.Á.F.-B.; funding acquisition, J.B.A.-H. and M.Á.F.-B. All authors have read and agreed to the published version of the manuscript.

Funding: Research supported by the grant PID2019-109099RB-C41 funded by MICIU/AEI/10.13039/501100011033 and by A way of making Europe (ERDF 2021-2027 Programmes) funded by the European Union FEDER program/funds.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethics Committee of University of Las Palmas de Gran Canaria (protocol code: CEIH-2014-01; date of approval: 17 July 2014).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data is not publicly available due to privacy restrictions as stated in resolution CEIH-2014-01 of the Ethics Committee of University of Las Palmas de Gran Canaria.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Molinuevo, L. Role of biomarkers in the early diagnosis of Alzheimer's disease. Rev. *Esp. Geriatr. Gerontol.* **2011**, *46* (Suppl. 1), 39–41.
2. Guix, J.L.M. Papel de los biomarcadores en el diagnóstico precoz de la enfermedad de Alzheimer. *Rev. Esp. Geriatr. Gerontol.* **2011**, *46*, 39–41. [CrossRef] [PubMed]
3. Andersen, C.K.; Wittrup-Jensen, K.U.; Lolk, A.; Andersen, K.; Kragh-Sørensen, P. Ability to perform activities of daily living is the main factor affecting quality of life in patients with dementia. *Health Qual. Life Outcomes* **2004**, *2*, 52. [CrossRef]
4. Association, A. 2017 Alzheimer's disease facts and figures. *Alzheimer's Dement.* **2017**, *13*, 325–373.
5. Laske, C.; Sohrabi, H.R.; Frost, S.M.; López-de-Ipiña, K.; Garrard, P.; Buscema, M. Innovative diagnostic tools for early detection of Alzheimer's disease. *Alzheimer's Dement.* **2015**, *11*, 561–578. [CrossRef]
6. Bäckman, L.; Jones, S.; Berger, A.-K.; Laukka, J.E.; Small, B.J. Cognitive impairment in preclinical Alzheimer's disease: A meta-analysis. *Neuropsychology* **2005**, *19*, 520–531. [CrossRef] [PubMed]
7. Deramecourt, V.; Lebert, F.; Debachy, B.; Mackowiak-Cordoliani, M.A.; Bombois, S.; Kerdraon, O.; Buée, L.; Maurage, C.-A.; Pasquier, F. Prediction of pathology in primary progressive language and speech disorders. *Neurology* **2010**, *74*, 42–49. [CrossRef] [PubMed]
8. McKhann, G.M.; Knopman, D.S.; Chertkow, H.; Hyman, B.T.; Jack, C.R., Jr.; Kawas, C.H.; Klunk, W.E.; Koroshetz, W.J.; Manly, J.J.; Mayeux, R.; 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.* **2011**, *7*, 263–269. [CrossRef]
9. Barragán-Pulido, M.L.; Alonso-Hernández, J.B.; Ferrer-Ballester, M.A.; Travieso-González, C.M.; Mekyska, J.; Smékal, Z. Alzheimer's disease and automatic speech analysis: A review. *Expert Syst. Appl.* **2020**, *150*, 113213. [CrossRef]
10. Kim, Y.; Lee, H.; Provost, E.M. Deep learning for robust feature generation in audiovisual emotion recognition. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013; pp. 3687–3691.
11. Khodabakhsh, A.; Yesil, F.; Guner, E.; Demiroglu, C. Evaluation of linguistic and prosodic features for detection of Alzheimer's disease in Turkish conversational speech. *EURASIP J. Audio Speech Music Process.* **2015**, *2015*, 9. [CrossRef]
12. Tanaka, H.; Adachi, H.; Ukita, N.; Kudo, T.; Nakamura, S. 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, Tokyo, Japan, 12–16 November 2016; pp. 261–265.
13. Rentoumi, V.; Paliouras, G.; Danasi, E.; Arfani, D.; Fragkopoulou, K.; Varlokosta, S.; Papadatos, S. Automatic detection of linguistic indicators as a means of early detection of Alzheimer's disease and of related dementias: A computational linguistics analysis. In Proceedings of the 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Debrecen, Hungary, 11–14 September 2017; pp. 33–38.

14. Winblad, B.; Amouyel, P.; Andrieu, S.; Ballard, C. Defeating Alzheimer's disease and other dementias: A priority for European science and society. *Lancet Neurol.* **2016**, *15*, 455–532. [CrossRef]
15. Farrús, M.; Codina-Filbà, J. Combining Prosodic, Voice Quality and Lexical Features to Automatically Detect Alzheimer's Disease. *arXiv* **2020**, arXiv:2011.09272v1.
16. Park, C.-Y.; Kim, M.; Shim, Y.; Ryoo, N.; Choi, H.; Jeong, H.T.; Yun, G.; Lee, H.; Kim, H.; Kim, S.; et al. Harnessing the Power of Voice: A Deep Neural Network Model for Alzheimer's Disease Detection. *Dement. Neurocogn. Disord.* **2024**, *23*, 1. [CrossRef] [PubMed]
17. Hajjar, I.; Okafor, M.; Choi, J.D.; Moore, E.; Abrol, A.; Calhoun, V.D.; Goldstein, F.C. Development of digital voice biomarkers and associations with cognition, cerebrospinal biomarkers, and neural representation in early Alzheimer's disease. *Alzheimer's Dement. Diagn. Assess. Dis. Monit.* **2023**, *15*, e12393. [CrossRef] [PubMed]
18. Liu, J.; Fu, F.; Li, L.; Yu, J.; Zhong, D.; Zhu, S.; Zhou, Y.; Liu, B.; Li, J. Efficient Pause Extraction and Encode Strategy for Alzheimer's Disease Detection Using Only Acoustic Features from Spontaneous Speech. *Brain Sci.* **2023**, *13*, 477. [CrossRef] [PubMed]
19. Campbell, E.L.; Mesía, R.Y.; Docío-Fernández, L.; García-Mateo, C. Paralinguistic and linguistic fluency features for Alzheimer's disease detection. *Comput. Speech Lang* **2021**, *68*, 101198. [CrossRef]
20. Cowie, R.; Cornelius, R.R. Describing the emotional states that are expressed in speech. *Speech Commun.* **2003**, *40*, 5–32. [CrossRef]
21. Chavhan, Y.D.; Yelure, B.S.; Tayade, K.N. Speech emotion recognition using RBF kernel of LIBSVM. In Proceedings of the 2015 2nd International Conference on Electronics and Communication Systems (ICECS), Coimbatore, India, 26–27 February 2015; pp. 1132–1135.
22. Balti, H.; Elmaghraby, A.S. Emotion analysis from speech using temporal contextual trajectories. In Proceedings of the 2014 IEEE Symposium on Computers and Communications (ISCC), Madeira, Portugal, 23–26 June 2014; pp. 1–7.
23. Laukka, P. Vocal Expression of Emotion: Discrete-Emotions and Dimensional Accounts. Ph.D. Thesis, Uppsala Universitet, Uppsala, Sweden, 2004.
24. Alonso, J.B.; Cabrera, J.; Medina, M.; Travieso, C.M. New approach in quantification of emotional intensity from the speech signal: Emotional temperature. *Expert Syst. Appl.* **2015**, *42*, 9554–9564. [CrossRef]
25. Goudbeek, M.; Scherer, K. Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *J. Acoust. Soc. Am.* **2010**, *128*, 1322. [CrossRef]
26. Kwon, O.W.; Chan, K.; Hao, J.; Lee, T.W. Emotion recognition by speech signals. In Proceedings of the Eighth European Conference on Speech Communication and Technology, Geneva, Switzerland, 1–4 September 2003.
27. Lee, C.; Narayanan, S. Emotion recognition using a data-driven fuzzy inference system. In Proceedings of the Eighth European Conference on Speech Communication and Technology, Geneva, Switzerland, 1–4 September 2003.
28. Harimi, A.; Shahzadi, A.; Ahmadyfard, A. Recognition of emotion using non-linear dynamics of speech. In Proceedings of the 7th International Symposium on Telecommunications (IST'2014), Tehran, Iran, 9–11 September 2014; pp. 446–451.
29. Altun, H.; Polat, G. Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection. *Expert Syst. Appl.* **2009**, *36*, 8197–8203. [CrossRef]
30. Amlerova, J.; Laczó, J.; Nedelska, Z.; Laczó, M.; Vyhnálek, M.; Zhang, B.; Sheardova, K.; Angelucci, F.; Andel, R. Emotional prosody recognition is impaired in Alzheimer's disease. *Alzheimer's Res. Ther.* **2022**, *14*, 50. [CrossRef] [PubMed]
31. Bhaduri, S.; Bhaduri, A.; Sarkar, R.; Analytics, M. Language Independent Speech Emotion and Non-invasive Early Detection of Neurocognitive Disorder. *arXiv* **2021**, arXiv:2106.01684v1.
32. Gong, Y.; Yang, L.; Zhang, J.; Chen, Z.; He, S.; Zhang, X.; Zhang, W. Using Speech Emotion Recognition as a Longitudinal Biomarker for Alzheimer's Disease. *Int. J. Biomed. Biol. Eng.* **2023**, *17*, 267–272. Available online: https://publications.waset.org/10013336/using-speech-emotion-recognition-as-a-longitudinal-biomarker-for-alzheimers-disease (accessed on 9 June 2024).
33. Bernieri, G.; Duarte, J.C. Identificação da Doença de Alzheimer Através da Fala Utilizando Reconhecimento de Emoções. *J. Health Informatics* **2023**, *15*, 1–14. [CrossRef]
34. López-De-Ipiña, K.; Alonso, J.B.; Solé-Casals, J.; Barroso, N.; Henriquez, P.; Faundez-Zanuy, M.; Travieso, C.M.; Ecay-Torres, M.; Martínez-Lage, P.; Eguiraun, H. On Automatic Diagnosis of Alzheimer's Disease Based on Spontaneous Speech Analysis and Emotional Temperature. *Cognit. Comput.* **2015**, *7*, 44–55. [CrossRef]
35. Lopez-de-Ipiña, K.; Alonso, J.B.; Barroso, N.; Faundez-Zanuy, M.; Ecay, M.; Solé-Casals, J.; Travieso, C.M.; Estanga, A. 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*; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7657, pp. 407–414.
36. López de Ipiña, K.; Alonso, J.B.; Solé-Casals, J.; Barroso, N.; Faundez, M.; Ecay, M.; Travieso, C.; Ezeiza, A.; Estanga, A. Alzheimer disease diagnosis based on automatic spontaneous speech analysis. In Proceedings of the IJCCI 2012: 4th International Joint Conference on Computational Intelligence, Barcelona, Spain, 5–7 October 2012; pp. 698–705.
37. López-De-Ipiña, K.; Alonso-Hernández, J.; Solé-Casals, J.; Travieso-González, C.; Ezeiza, A.; Faúndez-Zanuy, M.; Calvo, P.; Beitia, B. Feature selection for automatic analysis of emotional response based on nonlinear speech modeling suitable for diagnosis of Alzheimer's disease. *Neurocomputing* **2015**, *150*, 392–401. [CrossRef]
38. Barragán Pulido, M.L. Avances en el Análisis del Habla Mediante Sistemas Conversacionales Automáticos Aplicados a la Enfermedad de Alzheimer. Ph.D. Thesis, University of Murcia, Murcia, Spain, 2022.

39. Alonso-Hernández, J.B.; Barragán-Pulido, M.L.; Gil-Bordón, J.M.; Ferrer-Ballester, M.Á.; Travieso-González, C.M. Using a Human Interviewer or an Automatic Interviewer in the Evaluation of Patients with AD from Speech. *Appl. Sci.* **2021**, *11*, 3228. [CrossRef]
40. De Cheveigné, A.; Kawahara, H. YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.* **2002**, *111*, 1917–1930. [CrossRef] [PubMed]
41. Lazar, N.A. Basic Statistical Analysis. In *The Statistical Analysis of Functional MRI Data*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 1–36.
42. Sarhan, A.E. Estimation of the mean and standard deviation by order statistics. *Ann. Math. Stat.* **1954**, *25*, 317–328. [CrossRef]
43. Groeneveld, R.A.; Meeden, G. Measuring Skewness and Kurtosis. *J. R. Stat. Soc. Ser. D Stat.* **1984**, *33*, 391–399. [CrossRef]
44. Stata: Software for Statistics and Data Science. Available online: https://www.stata.com/ (accessed on 20 November 2019).
45. GoodData. Normality Testing-Skewness and Kurtosis. 2020. Available online: https://www.gooddata.com/ (accessed on 18 January 2022).
46. Wilcoxon, F. Some rapid approximate statistical procedures. *Ann. N. Y. Acad. Sci.* **1950**, *52*, 808–814. [CrossRef]
47. Kruskal, W.H.; Wallis, W.A. Use of Ranks in One-Criterion Variance Analysis. *J. Am. Stat. Assoc.* **1952**, *47*, 583–621. [CrossRef]
48. Hernández, J.; Pulido, M.; Bordón, J.; Ballester, M.; González, C. Speech Evaluation of patients with Alzheimer's Disease using an automatic interviewer. *Expert Syst. Appl.* **2022**, *192*, 116386. [CrossRef]