INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Automatic Assessment of Alzheimer's across Three Languages Using Speech and Language Features

P. A. Pérez-Toro[1,2*], *T. Arias-Vergara*[1,2,3], *F. Braun*[4], *F. Hönig*[5], *C. A. Tobón-Quintero*[6], *D. Aguillón*[6], *F. Lopera*[6], *L. Hincapié-Henao*[6], *M. Schuster*[3], *K. Riedhammer*[4], *A. Maier*[1] *E. Nöth*[1], *J. R. Orozco-Arroyave*[1,2]

[1]Pattern Recognition Lab. Friedrich-Alexander Universität, Erlangen-Nürnberg, Erlangen, Germany
[2]GITA Lab, Facultad de Ingeniería. Universidad de Antioquia, Medellín, Colombia
[3]Department of Otorhinolaryngology, Head and Neck Surgery. Ludwig-Maximilians University, Munich, Germany
[4]Technische Hochschule Nürnberg, Nuremberg, Germany
[5]KST Institut GmbH, Bad Emstal, Germany
[6] Neuroscience Research Group, Medical Research Institute, Faculty of Medicine, Universidad de Antioquia, Medellín, Colombia

*paula.andrea.perez@fau.de

Abstract

With the increasing prevalence of Alzheimer's Disease (AD) worldwide, it is essential to develop non-invasive methods to monitor the progression of the disease. Speech and language analyses are suitable for detecting the cognitive impairment of AD patients; thus, by analyzing changes in speech patterns and language use, researchers can develop methods to monitor AD remotely. In this paper, we investigated several speech and language techniques commonly used for the automatic detection of AD. Furthermore, we considered speech recordings of 448 patients in three different languages: Spanish (57), German (205), and English (186). Cross-lingual analysis was carried out using two classification approaches: (1) training/testing in one or more languages and (2) training in one language and testing in another. We obtained unweighted average recall values of up to 83% to classify AD using the first classification approach and up to 70% with the second.

Index Terms: Pathological Speech Processing, Alzheimer's Disease, Dementia, Deep Learning, Cross-lingual Analysis

1. Introduction

Alzheimer's Disease (AD) is a common form of dementia and the most prevalent neurodegenerative disorder. It is characterized by changes in brain tissue and a reduction in the neurotransmitter acetylcholine, which can have detrimental effects on memory, language, comprehension, and behavior, leading to significant impairments in language and communication [1]. This global trend of demographic aging reflects advances in healthcare, resulting in people living longer and healthier lives. As a result, there is an increasing proportion of older individuals worldwide. While dementia primarily affects older adults, there is a growing recognition of cases that begin before age 65 [2]. It is often caused by genetic mutations, for instance, the PSEN1-E280A or "Paisa mutation" which includes Early-Onset Alzheimer's (EOA), typically diagnosed at a mean age of 49 years [3]. For this reason, it is important to develop speech technology for the detection and monitoring of the disease that in the future can lead to an early diagnosis.

1.1. Related work

Previous research has investigated the use of automatic speech and language analysis to diagnose and assess AD. Traditionally, prosodic measures in dementia research have focused on temporal factors, intensity, voice quality, voice periods, and variation in fundamental frequency (F_0). In addition to these features, other acoustic features such as formant frequencies, Mel-Frequency Cepstral Coefficients (MFCCs), and Energy distributed in the Bark scale (BBE) can provide contextualized interpretations of the acoustic information. More recently, researchers have explored using neural embeddings, such as Wav2Vec, x-vector, and i-vectors, to assess AD, aiming to capture relevant information about the speaker in a condensed format [4, 5, 6, 7].

Most studies use English datasets, such as the Pitt Corpus from the Dementia Bank, comprised solely of native American English speakers. As a result, numerous studies have used this corpus to investigate AD. Although significant research has been conducted using this dataset, including two Interspeech and one ICASSP Challenge, only some studies have used non-English data [8, 9, 10, 6, 11, 12]. This represents a considerable gap in the field, which could lead to global inequity in dementia research. For instance, the Hungarian MCI-mAD database [13], Mandarin_Lu corpus [14], and a Chilean-Spanish AD [14] corpus has been used in studies that have reported accuracies of up to 86%.

Cross-lingual studies have a significant gap in dementia research, as they can lead to disparities between different studies. A transfer learning strategy using English (Pitt corpus) and Spanish (Chilean-Spanish) data achieved good performance in [15]. The authors reported Unweighted Average Recalls (UAR) of up to 85% while combining the corpora seems more challenging (UAR=66%). In [16], a study based on classical speech and language features explored the feasibility of cross-linguistic AD detection in English and German. The authors reported that separate training in each corpus achieved good performance (German UAR=86%, English UAR=77%). The authors reported results slightly above chance training and testing in different languages.

1.2. Contribution of this work

In this paper, we investigated the suitability of speech and language analyses for the automatic detection of dementia in AD using a cross-lingual approach with three different languages: English (EN), Spanish (ES), and German (DE). Cross-lingual analysis was carried out using two classification approaches: (1) training/testing in one or more languages and (2) training in one language and testing in another. For this, we considered speech recordings of 205 German native speakers, 186 English native speakers, and 57 Colombian native speakers. We used Support Vector Machine (SVM) and Artificial Neural Network (ANN) for classification. Acoustic and linguistic features were considered for modeling the signals. In the case of acoustic, we computed duration, rhythm, pleasure arousal dominance, and Wav2Vec embeddings. For linguistics, we computed word embeddings and grammar features. We also proposed a fusion strategy for different feature sets using the ANN, which consists of concatenating the features as "channels" in the input tensor, resulting in better results than using an early fusion strategy.

10.21437/Interspeech.2023-2079

2. Data

We used three datasets containing speech recordings and transcripts in English, Spanish, and German for AD detection. The datasets included semi-spontaneous speech recordings [17] and transcripts from picture description tasks. Furthermore, the interviewer's speech was removed from the recordings based on the timestamps provided in the datasets.

2.1. English Dataset

For this study, a subset of the Pitt Corpus [18] was used, comprising a total of 186 native English speakers. The group with Alzheimer's Disease (AD) was age- and sex-matched with the Healthy Control (HC) group. The speech task consisted of recordings [17] and manual transcripts describing the "cookie theft picture" [19].

2.2. Spanish Dataset

The dataset [20] includes recordings of spontaneous speech and transliterations from 57 Spanish speakers from Colombia, comprising 8 EOA patients with Mild Dementia (MD), 23 EOA patients with Mild Cognitive Impairment (MCI), and 27 HC subjects. The patients are genetic carriers of the "Paisa mutation" [3]. The participants were asked to describe the cookie theft picture, as in the English dataset. The transcripts were created manually.

2.3. German Dataset

This corpus comprises 205 native German speakers (109 female, 96 men), which was recorded in a multi-site study using a uniform digital tablet platform and was provided by the PARLO Institute for Research and Teaching in Speech Therapy[1]. The dataset consists of 83 patients with MCI, 83 with dementia, and 59 HC subjects. Unlike the English and Spanish dataset, the participants describe a different picture, which is one of the tasks in this dataset. For this dataset, the transcriptions were automatically generated using Whisper [21].

Patients with MCI, MD, and dementia were grouped together for the class AD. Additional demographic information from the three datasets is displayed in Table 1.

Table 1: *Demographic information of the subjects for each language*

	AD Patients F/M	HC Subjects F/M
	English Corpus	
Number of Subjects	33/60	37/56
Age [years]	66.5 (7.8)/70.0 (7.3)	64.5 (8.1)/63.3 (8.0)
	Spanish Corpus	
Number of Subjects	14/15	15/12
Age [years]	48.2 (5.7)/50.7 (7.1)	49.5 (7.7)/ 53.2 (7.1)
	German Corpus	
Number of Subjects	83/64	26/32
Age [years]	70.1 (8.4)/70.6 (8.2)	68.6 (8.5)/ 72.6 (7.8)

Values are expressed as mean (standard deviation). F: female. M: male. Age is given in years.

3. Methods

3.1. Linguistic Features

Grammar Based: We considered the features previously used for automatic AD detection [22]. They aimed at modeling the sentence structuring capabilities of AD patients, who exhibit deficits in using nouns and verbs and difficulties using verbs when arguments are involved [23]. Our goal is to evaluate their sentence structuring capabilities by counting the elements involved in structuring sentences, as well as the number of grammatical elements, such as verbs and nouns, found in their transcripts. We considered Part-Of-Speech (POS) counts, namely the Noun to Verb Ratio (NVR), Noun Ratio (NR), Pronoun Ratio (PR), and Subordinated Coordinated Conjunctions Ratio (SCCR). These POS counts assess the syntactical abilities of AD patients in structuring sentences. This set of features also includes the Flesch reading score, Flesch-Kincaid grade level, propositional density, and content density of the transcript [24].

Word Embeddings: This study considers word embeddings, specifically Bidirectional Encoder Representations from Transformers (BERT) [25]. These methods form direct connections between individual elements through a process called attention and use transfer learning. BERT uses various attention mechanisms known as heads, which function simultaneously and capture a broader range of relationships between words through multi-head attention. The model is trained using transfer learning, where it is initially trained on two unsupervised tasks. The first task involves Masked Language Modeling (MLM), where the system predicts missing words (masked) in a sentence. The second task is Next Sentence Prediction (NSP), where the model predicts if a sentence follows another. The last layer (768 units) is taken as the word-embedding representation, and the mean of overall word-embeddings is computed for the classification task. The model was trained with the Wikicorpus data from 102 languages for BERT-Base. The source code is also available online[2] [26].

3.2. Acoustic Features

Duration: Previous studies on AD also used these features and achieved satisfactory results [11]. The feature set comprises duration-based descriptors that were obtained using an energy-based Voice Activity Detection (VAD) algorithm, which identifies the speech and pause segments. The descriptors include the count of pauses and speech segments per second, the ratio of speech segments to pauses, and six functionals (mean, standard deviation, kurtosis, skewness, minimum, and maximum) that characterize the duration of pauses and speech segments.

Rhythm: We used the metrics proposed in [27] and [28] to model timing information extracted from vowels and consonants. Three main descriptors are considered: (1) the raw and normalized Pairwise Variability Index (rPVI and nPVI, respectively) to measure the duration variability of successive vowel and consonant intervals, (2) the Global Proportions of Intervals (GPI) to measure the vowels produced per second, and (3) the standard deviation of the vowel/consonant (dGPI) duration intervals. To detect the consonants and vowels, we trained three phoneme recognizers in Spanish, German, and English. The model was implemented using recurrent neural networks with LSTM cells. Regardless of the language, the network ar-

[1] https://www.parlo-institut.de/

[2] https://github.com/PauPerezT/WEBERT

chitecture consists of two convolution layers to process Mel-spectrograms (50×64) with ReLU activation functions, two max-pooling layers, and dropout. The feature maps are concatenated to form the sequence of feature vectors processed by two stacked bidirectional LSTMs with 512 hidden units. Then, a sigmoid activation function is used to predict the sequence of phonemes. For English, the network was trained with the TIMIT corpus [29]; for German, we used Verbmobil [30]; and for Spanish, we used the TEDx Spanish Corpus [3].

Pleasure Arousal Dominance (PAD): We proposed these features in a previous study addressing the automatic detection of depression detection in Parkinson's disease [31, 11]. It consists of a pre-trained deep neural network designed to extract emotional information based on three dimensions: (1) valence, which represents the pleasantness or unpleasantness of an emotion; (2) arousal, which represents the level of activation or agitation; and (3) dominance, which represents the level of control or influence of an individual over a situation [32]. Three models were trained on the IEMOCAP database [33], addressing three binary-classification problems: active vs. passive arousal, positive vs. negative valence, and strong vs. weak dominance. The model takes a 3D-multi-channel log-magnitude Mel spectrogram as input, with each dimension formed by a sequence of 500 ms and three different resolution windows. It combines CNN and GRU to model different articulation and prosody information aspects. The final linear layer consists of 2 units, which use a sigmoid activation function to obtain the posterior probabilities and independently observe each dimension's contribution. Four functionals are computed across the sequences to form a 24-dimensional static vector. The source code is available online[4].

Acoustic Embeddings: These embeddings are generated using the Wav2Vec 2.0 model, which employs self-supervised learning methods to learn representations from raw speech signals. The model consists of three main components: feature extraction, a context network, and a linear projection to the output. The input consists of 16 kHz raw audio split into 25 ms chunks with 10 ms frame shift. The feature extraction part uses temporal convolutions to convert speech information into a latent space representation. Similar to BERT, audio segments are masked and quantized for self-supervised training, and a contextualized representation is obtained through a Transformer-based approach with contrastive learning. This method has been fine-tuned for emotion recognition, speaker verification, and speech disorder assessment tasks. We focus on the pre-trained Wav2Vec XLSR-53 and experiment with different layers, specifically the latent layer, the twelve layers within the attention mechanism, and the last layer. Our results indicate that the model's first, ninth, and twelfth layers were the most effective for our experiments. We compute the mean value over the output chunks for each layer, resulting in a fixed vector of 768 elements per speaker and layer.

3.3. Automatic Detection of AD

Two different classifiers were considered for comparison purposes: an SVM and an ANN. The classifiers were optimized following a nested 4-fold cross-validation strategy. The performance of the classifiers was measured in terms of UAR, sensitivity, and specificity. In the case of the SVM, the optimal parameters were found through a grid search, where $C \in \{10^{-4}, 10^{-3}, ..., 10^{4}\}$ and $\gamma \in \{10^{-4}, 10^{-3}, ..., 10^{4}\}$. For the ANN, we proposed a light fusion architecture to combine the different feature sets and reduce the risk of overfitting due to the size of the datasets. Figure 1 shows the proposed model. For the network input, we concatenate the feature sets as "channels" in a tensor of dimensions $1 \times 768 \times C$, where C is the number of feature set combinations and 768 is the maximum number of features in one set, which in our case are from the word embeddings. Additionally, we use padding on the feature sets with less than 768 elements. The 1-Dimensional convolutional layer aims to fuse the feature sets at an intermediate level (middle fusion), followed by a pooling layer with an ELU activation, a linear layer of 256 units, and the classification layer.

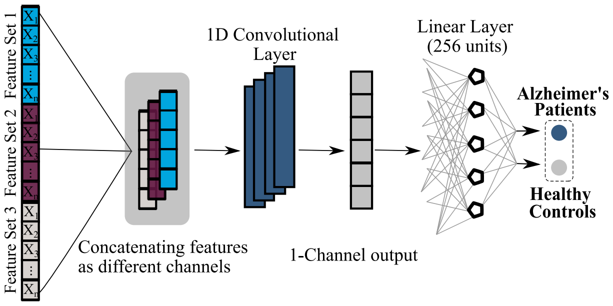

Figure 1: *Architecture of the ANN proposed to perform classification and feature fusion.*

4. Experiments and Results

We performed experiments taking into account all the combinations between languages with the aim of observing if it is feasible to combine the different corpora from different language sources. We used two classification approaches for language-dependent and cross-lingual analyses: (1) training/testing in one or more languages and (2) training in one language and testing in another.

Table 2: *Best classification results obtained for each language and possible combinations.* ***UAR:*** *Unweighted Average Recall.* ***Sens:*** *Sensitivity.* ***Spe:*** *Specificity.*

Language	Feature Fusion	Classifier	UAR	Sens	Spe
EN	WV1 + WV12	SVM	70	58	82
	BERT + WV1	ANN	**82**	**89**	**74**
ES	PAD + Rhythm	SVM	75	79	70
	Gr + WV12	ANN	**78**	**76**	**80**
DE	Rhythm + WV1 + WV9 + WV12	SVM	65	55	74
	BERT + WV1 + WV9	ANN	**70**	**79**	**62**
EN+ES	PAD + Dur + WV1 + WV12 + WV9	SVM	72	64	79
	BERT + Rhythm + Gr + WV12	ANN	**78**	**70**	**86**
EN+DE	Rhythm	SVM	55	44	66
	BERT + PAD + Rhythm + WV1 + WV9	ANN	67	65	70
ES+DE	PAD + WV9	SVM	66	65	67
	Dur + PAD + Rhythm + WV1 + WV9	ANN	68	72	64
EN + ES + DE	Rhythm	SVM	55	45	65
	Rhythm + WV1 + WV9	ANN	**73**	**81**	**66**

WVi: Wav2Vec i-th layer of the transformer. PAD: Pleasure Arousal Dominance posteriors. Dur: Duration features. Gr: Grammar features.

The results for the first classification approach are shown

[3] http://www.ciempiess.org/downloads

[4] https://github.com/PauPerezT/PADS

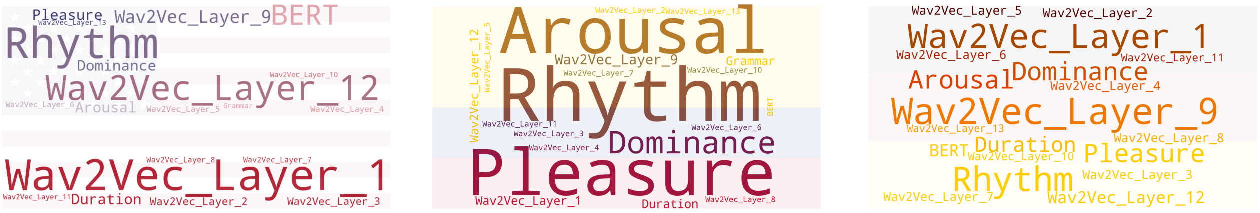

Figure 2: *Feature importance for each language: English (left), Spanish (middle), and German (right).*

in Table 2, where only the best combinations of features are reported. In particular, we achieved an UAR of 82% in English, 78% in Spanish, and 70% in German. When combining two languages, the best result was obtained when training English and Spanish together (UAR= 78%). For the three languages, the classification performance was 73%, which is not an improvement on the other languages due to the German speakers. Regarding the linguistic and acoustic feature sets, we performed a word cloud analysis to visualize the most important features resulting in the best classification performance. The word clouds are shown in Figure 2.

The most frequent features across the three languages are related to acoustic information, particularly the rhythm-based features, which are used to measure the variability in the duration of consecutive vowels (and consonants). For each language, other important features are: for English, the acoustic embeddings (the first, ninth, and twelve layers from Wav2Vec) and word embeddings (BERT); for Spanish, our emotional embeddings; and for German, a combination of duration features, acoustic, word, and emotional embeddings. Table 3 shows the classification results obtained when the classifiers were trained with one language and tested with another. The performance of the classifiers is considerably lower than those presented in Table 2. We obtained comparable results when we trained an ANN in German and tested it in Spanish, which shows that features like Rhythm and Dur-based are "transferable" when another language is also included in the training. These results are expected if we consider that rhythm features exhibit language-dependent patterns. For instance, the vocalic variability in Spanish (measured with the PVI) has been shown to be lower than in German [27] due to the language structure, e.g., the Spanish language does not differentiate between long or short vowels, but German and English do.

Table 3: *Best classification results obtained while training in one language and testing in another.* ***UAR:*** *Unweighted Average Recall.* ***Sens:*** *Sensitivity.* ***Spe:*** *Specificity.*

Train	Test	Feature Fusion	Classifier	UAR	Sens	Spe
EN	ES	BERT+WV1	SVM	60	34	85
		PAD+Dur+Rhythm+WV1+WV9	ANN	65	70	59
	DE	PAD+Rhythm	SVM	55	45	66
		PAD+Rhythm	ANN	57	26	88
ES	EN	PAD+Dur+WV9+WV12	SVM	67	68	67
		WV1	ANN	64	74	54
	DE	Dur+Rhythm+WV1+WV12	SVM	56	27	84
		BERT+WV9	ANN	59	50	67
DE	EN	PAD+Dur+WV1+WV9+WV12	SVM	62	32	92
		Dur+Rhythm	ANN	64	69	60
	ES	WV12	SVM	56	31	81
		Dur+Rhythm	ANN	**70**	**78**	**62**

WVi: Wav2Vec i-th layer of the transformer. PAD: Pleasure Arousal Dominance posteriors. Dur: Duration features. Gr: Grammar features.

5. Discussion and Conclusion

In this paper, we investigated the suitability of speech and language analyses for detecting AD automatically. For this, we considered speech recordings in Spanish, English, and German, which we modeled with several combinations of features. Based on the analysis made in a previous study from last year's Interspeech [15], multilingual word embeddings are influenced by language-specific characteristics (English and Spanish), in consequence, this paper aims to assess the transferability of features between languages when trained in one and tested in another, contrary to relying on transfer learning techniques.

The classification task was performed using two training/test approaches: (1) training and testing in the same language or combination of languages and (2) training in one language and testing in another. The results obtained with the first approach showed that the best classification results were obtained when automatic detection of AD was performed for each language. Overall, German speakers were the most difficult to classify, and combining them with other languages did not improve the results. This might be due to the mismatch between the speech task performed by the Spanish/English speakers (cookie theft picture description) and the German speakers (picture description other than cookie theft). One limitation is the age difference between the participants in other languages compared to Spanish. Another problem is the fact that MCI is considered to belong to the class AD. Most of the studies use a Mini-mental State Examination (MMSE) result of $\leq$ 24 as an indicator of AD. Yet, it could also be MCI patients. In future work, we will look at a 3-class problem for classifying between HC, MCI, and AD.

Regarding the combination of features, the rhythm was the most relevant feature set for classifying AD in each language; however, the results obtained with the second training approach showed that the information provided by the features is not easily transferable across languages. Some features, such as rhythm-based, can differentiate between patients and controls in individual languages; however, it is necessary to introduce information about other languages during training, so the model learns to "focus" on differences due to the pathology.

In the future, we will explore individual features in detail. We expect a variation of the F_0 to be transferable, while e.g., variation in vowels commonly is more difficult.

6. Acknowledgments

This study was partially funded by the University of Antioquia grant # PI2023-58010.

7. References

[1] D. S. Knopman, H. Amieva, R. C. Petersen, G. Chételat, D. M. Holtzman, B. T. Hyman, R. A. Nixon, and D. T. Jones, "Alzheimer disease," *Nature reviews Disease primers*, vol. 7, no. 1, pp. 1–21, 2021.

[2] W. H. Organization, *World report on ageing and health*. World Health Organization, 2015.

[3] M. Lalli *et al.*, "Origin of the PSEN1 E280A mutation causing early-onset Alzheimer's disease," *Alzheimer's & Dementia*, vol. 10, pp. S277–S283, 2014.

[4] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

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

[6] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To BERT or Not To BERT: Comparing Speech and Language-based Approaches for Alzheimer's Disease Detection," *Proc. Interspeech 2020*, 2020.

[7] I. Vigo, L. Coelho, and S. Reis, "Speech-and language-based classification of alzheimer's disease: A systematic review," *Bioengineering*, vol. 9, no. 1, p. 27, 2022.

[8] M. Martinc and S. Pollak, "Tackling the ADReSS challenge: a multimodal approach to the automated recognition of Alzheimer's dementia," *Proc. Interspeech 2020*, pp. 2157–2161, 2020.

[9] R. Pappagari, J. Cho, L. Moro-Velazquez, and N. Dehak, "Using state of the art speaker recognition and natural language processing technologies to detect Alzheimer's disease and assess its severity," *Proc. Interspeech 2020*, pp. 2177–2181, 2020.

[10] M. S. S. Syed, M. Lech, and E. Pirogova, "Automated Screening for Alzheimer's Dementia through Spontaneous Speech," *Proc. Interspeech 2020*, pp. 1–5, 2020.

[11] P. A. Pérez-Toro, S. P. Bayerl, T. Arias-Vergara, J. C. Vásquez-Correa, P. Klumpp, M. Schuster, E. Nöth, J. R. Orozco-Arroyave, and K. Riedhammer, "Influence of the Interviewer on the Automatic Assessment of Alzheimer's Disease in the Context of the ADReSSo Challenge." in *Interspeech*, 2021, pp. 3785–3789.

[12] Z. Ye, S. Hu, J. Li, X. Xie, M. Geng, J. Yu, J. Xu, B. Xue, S. Liu, X. Liu *et al.*, "Development of the cuhk elderly speech recognition system for neurocognitive disorder detection using the dementiabank corpus," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6433–6437.

[13] G. Gosztolya, V. Vincze, L. Tóth, M. Pákáski, J. Kálmán, and I. Hoffmann, "Identifying Mild Cognitive Impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features," *Computer Speech & Language*, vol. 53, pp. 181–197, 2019.

[14] Y.-W. Chien, S.-Y. Hong, W.-T. Cheah, L.-H. Yao, Y.-L. Chang, and L.-C. Fu, "An automatic assessment system for Alzheimer's disease based on speech using feature sequence generator and recurrent neural network," *Scientific Reports*, vol. 9, no. 1, pp. 1–10, 2019.

[15] P. A. Pérez-Toro, P. Klumpp, A. Hernandez, T. Arias-Vergara, P. Lillo, A. Slachevsky, A. M. Garcıa, M. Schuster, A. K. Maier, E. Noeth *et al.*, "Alzheimer's detection from English to Spanish using acoustic and linguistic embeddings," in *inProc. Annu. Conf. Int. Speech Commun. Assoc*, 2022, pp. 2483–2487.

[16] A. Ablimit, C. Botelho, A. Abad, T. Schultz, and I. Trancoso, "Exploring Dementia Detection from Speech: Cross Corpus Analysis," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6472–6476.

[17] V. Boschi, E. Catricala, M. Consonni *et al.*, "Connected speech in neurodegenerative language disorders: a review," *Frontiers in psychology*, vol. 8, p. 269, 2017.

[18] 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," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[19] H. Goodglass *et al.*, "Cookie Theft picture," *Boston diagnostic aphasia examination. Philadelphia, PA: Lea & Febiger*, 1983.

[20] P. A. Pérez-Toro, J. C. Vásquez-Correa, T. Arias-Vergara, P. Klumpp, M. Sierra-Castrillón, M. E. Roldán-López, D. Aguillón, L. Hincapié-Henao, C. A. Tobón-Quintero, T. Bocklet, M. Schuster, J. R. Orozco-Arroyave, and E. Nöth, "Acoustic and Linguistic Analyses to Assess Early-Onset and Genetic Alzheimer's Disease," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 8338–8342.

[21] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[22] J. S. Guerrero-Cristancho, J. C. Vásquez-Correa, and J. R. Orozco-Arroyave, "Word-embeddings and grammar features to detect language disorders in alzheimer's disease patients," *TecnoLógicas*, vol. 23, no. 47, pp. 63–75, 2020.

[23] M. Kim and C. K. Thompson, "Verb deficits in Alzheimer's disease and agrammatism: Implications for lexical organization," *Brain and language*, vol. 88, no. 1, pp. 1–20, 2004.

[24] J. N. Farr, J. J. Jenkins, and D. G. Paterson, "Simplification of Flesch reading ease formula." *Journal of applied psychology*, vol. 35, no. 5, p. 333, 1951.

[25] J. Devlin, M. W. Chang, K. Lee *et al.*, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[26] P. A. Perez-Toro, "PauPerezT/WEBERT: Word Embeddings using BERT," https://doi.org/10.5281/zenodo.3964244, Jul. 2020.

[27] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," *Papers in laboratory phonology*, vol. 7, no. 515-546, 2002.

[28] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 73, no. 3, pp. 265–292, 1999.

[29] J. S. Garofolo, L. Lamel, W. M. Fisher *et al.*, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[30] W. Wahlster, *Verbmobil: foundations of speech-to-speech translation*. Springer Science & Business Media, 2013.

[31] P. A. Pérez-Toro, J. C. Vasquez-Correa, T. Arias-Vergara, P. Klumpp, M. Schuster, and J. R. Nöth, E.and Orozco-Arroyave, "Emotional state modeling for the assessment of depression in Parkinson's disease," in *Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings 24*. Springer, 2021, pp. 457–468.

[32] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, pp. 261–292, 1996.

[33] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.