Alzheimer’s Detection from English to Spanish Using Acoustic and Linguistic Embeddings

P. A. Pérez-Toro1,2 ⋆, P. Klumpp1, A. Hernandez1, T. Arias-Vergara1,2,7, P. Lillo3, A. Slacheksky3, A. M. Garcia1,5,6, M. Schuster7, A. K. Maier1, E. Nöth4, J. R. Orozco-Arroyave1,2.

1Pattern Recognition Lab. Friedrich-Alexander Universität, Erlangen-Nürnberg, Erlangen, Germany
2GITA Lab, Facultad de Ingeniería, Universidad de Antioquia, Medellín, Colombia
3Departamento de Neurología Sur, Facultad de Medicina, Universidad de Chile, Santiago, Chile
4Centro de Neurociencias Cognitivas, Universidad de San Andrés, Vito Dumas 284, Buenos Aires, Argentina
5Global Brain Health Institute, University of California, San Francisco, US
6Departamento de Lenguística y Literatura, Facultad de Humanidades, Universidad de Santiago de Chile, Santiago, Chile
7Department of Otalaryngology, Head and Neck Surgery, Ludwig-Maximilians University, Munich, Germany

⋆corresponding authors: paula.andrea.perez@fau.de

Abstract

Cross-lingual approaches are growing in popularity in the machine learning domain, where large amounts of data are required to obtain better generalizations. Moreover, one of the biggest problems is the availability of clinical speech data, where most of the resources are in English. For instance, not many available Alzheimer’s Disease (AD) corpora in different languages can be found in the literature. Despite the phonological and phonemic differences between Spanish and English, fortunately, there are also similarities between these two languages, e.g., around 40% of all words in English have a related word in Spanish. In this work, we want to investigate the feasibility of combining information from English and Spanish languages to discriminate AD. Two datasets were considered: part of the Pitt Corpus, which is composed of English speakers, and a Spanish AD dataset composed of speakers from Chile. We based our analysis on known acoustic (Wav2Vec) and word (BERT, RoBERTa) embeddings using different classifiers. Strong language dependencies were found, even using multilingual representations. We observed that linguistic information was more important for classifying English AD (F-Score=0.76) and acoustic for Spanish AD (F-Score=0.80). Using knowledge transferred from English to Spanish achieved F-scores of up to 0.85 for discriminating AD.

Index Terms: Alzheimer’s Disease, Cross-Lingual Analysis, Acoustic Embeddings, Linguistic Embeddings

1. Introduction

The most common type of dementia is Alzheimer’s Disease (AD), which is characterized by progressive cognitive decline. It is caused by brain tissue changes and a loss of a chemical vital to brain function called acetylcholine [1]. This adversely affects memory, language, understanding, and behavior. Thus, language and para-linguistic aspects can be highly compromised, triggering problems related to the capability to produce coherent language.

Clinical assessments draw heavily on neuropsychological tests, such as the Mini-Mental State Examination [2], which is a 30-point scale, where classically scores of over 24 indicates normal cognition. However, this test proves suboptimally reliable [3, 4, 5] and requires specialized examiners who are vastly outnumbered by current and prospective patients, calling for new diagnostic-support tools.

Automatic speech and language analyses have been considered in the literature for supporting the diagnosis and evaluation of AD and a majority of them have been conducted only using English corpora. One of the best-known Alzheimer’s speech datasets available is the Pitt Corpus from the Dementia Bank [6], which is exclusively composed of native American English speakers. Since this dataset is publicly available, a large prior work has been reported in the literature using this corpus [7, 8, 9, 10, 11, 12], including two Interspeech Challenges [13, 14]. Moreover, there are some studies that used English cross-corpus learning for detecting Alzheimer’s [15].

Conversely, very few AD studies have used non-English data [16, 17, 18]. Some examples include: (1) the “Hungarian MCI-mAD database” that was used in [16], where accuracies of up to 86% were obtained, (2) the “Mandarin_Lu corpus” from the Dementia Bank in [17], which reported an area under the receiver operating characteristic curve of up to 0.90, and (3) a Chilean-Spanish AD corpus in [18], which achieved accuracies of up to 71%.

Cross-lingual studies are a major gap in the field, potentially representing a new source of global inequity in dementia research. Dementia has a prevalence of 7.1% in Latin America [19], where the risk factors explain 56% of dementia cases [20]. The Latin American population is made up mostly of native Spanish speakers. Spanish is the second most spoken language in the world only behind Mandarin Chinese and above English [21]. English and Spanish exhibit main-fold differences, crucially including phonological and lexical distinctions. One of the most noticeable differences is that excluding diphthongs, English has fifteen vowels whereas Spanish has five. Whereas there are differences between these two languages, there are also some similarities such as the use of the Roman alphabet, 40% of all English words have cognates in Spanish, learning uses similar basic processes, and despite some exceptions related to adjectives, there are similarities in sentence structures [22].

The reasons mentioned above motivated the use of multilingual techniques to transfer information from English to Spanish in this study. A set of fine-tuned multilingual acoustic and linguistic embeddings based on “Transformers” [23] is considered to perform transfer learning, since in the literature, these methods have shown good results on similar tasks [7, 10, 8, 24]. We explored these approaches on cross-lingual data because most studies performed experiments only on one language. Our findings show that by using transfer learning from one language
to another the classification accuracy is improved. As expected, linguistic information was strongly linked to the language structure when transferring information from English to Spanish. However, the acoustic embeddings were more suitable for transferring knowledge for classifying Spanish AD.

2. Methods

2.1. Data

We employed datasets previously reported for English [6] and Spanish [18]. All the recordings were denoised using the proposed model in [25], which is based on recurrent networks that uses complex linear coding to estimate a denoising mask.

2.1.1. The Pitt Corpus

A sub-set with a total of 186 speakers from the Pitt Corpus [6] was considered in this study. The participants are native English speakers, and the group with AD was matched by age and sex with respect to the group of Healthy Control (HC) individuals. The speech task consisted of semi-spontaneous speech recordings [26] and transcripts describing the “cookie theft picture” [27]. The speech of the interviewer was removed according to the timestamps provided by the dataset. After segmentation, the average duration was 87±23 seconds for the AD patients and 65±23 seconds for the HC subjects. Demographic information of the participants and the MMSE are provided in Table 1.

Table 1: Demographic and clinical information of the subjects in the subset of the Pitt corpus

<table>
<thead>
<tr>
<th></th>
<th>AD patients</th>
<th>HC subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of subjects</td>
<td>33/36</td>
<td>37/36</td>
</tr>
<tr>
<td>Age [years]</td>
<td>66.5 (7.8) /70.0 (7.3)</td>
<td>64.5 (8.1) /63.3 (8.0)</td>
</tr>
<tr>
<td>Education [years]</td>
<td>13.5 (2.8) /11.8 (2.2)</td>
<td>13.7 (2.5) /14.1 (2.6)</td>
</tr>
<tr>
<td>MMSE</td>
<td>20.0 (5.5) /19.3 (4.9)</td>
<td>28.8 (1.1) /29.3 (1.0)</td>
</tr>
</tbody>
</table>

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

2.1.2. Chilean Spanish AD

This dataset was recorded by the Universidad de Chile and Hospital del Salvador (Chile) [18]. The data consist of semi-spontaneous speech recordings and their transcriptions from 39 Chilean Spanish speakers (21 AD and 18 HC) while describing the “cookie theft picture” [27]. Participants were matched by sex, age, and education. They were evaluated by an expert neurologist following the National Institute of Neurological and Communicative Diseases and Stroke-Alzheimer’s Disease and Related Disorders Association clinical criteria for AD. The recordings were manually segmented and transcribed. After removing the interviewer’s parts, the average duration of the recordings is 84±29 seconds for the AD patients and 73±15 seconds for the HC subjects. The demographic information of the participants is provided in Table 2.

Table 2: Demographic and clinical information of the subjects in the Chilean Spanish AD corpus

<table>
<thead>
<tr>
<th></th>
<th>AD patients</th>
<th>HC subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of subjects</td>
<td>13/14</td>
<td>17/4</td>
</tr>
<tr>
<td>Age [years]</td>
<td>79.1 (6.8) /77.0 (7.1)</td>
<td>74.5 (1.7) /76.4 (4.6)</td>
</tr>
<tr>
<td>Education [years]</td>
<td>11.1 (3.9) /11.4 (3.7)</td>
<td>7.5 (3.6) /14.4 (3.3)</td>
</tr>
<tr>
<td>MMSE</td>
<td>20.7 (4.9) /21.9 (3.1)</td>
<td>--</td>
</tr>
</tbody>
</table>

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

2.2. Pre-trained Embeddings

A set of multilingual pre-trained embeddings based on “Transformers” [23] was considered. The main idea of these models is that the speech and language characteristics common in several languages can be learned at the same time. Thus, a compressed representation obtained from speech and text will contain transferable contextual knowledge to be applied in different languages.

2.2.1. Linguistic Embeddings

The methods considered here are based on the “word-embeddings” paradigm, which aims to capture semantic and syntactic relationships between words as well as contextual information. In this study, we considered the use of two pre-trained word-embeddings based on the encoder from the “Transformers” architecture [23], namely: Bidirectional Encoder Representations from Transformers (BERT) [28] and A Robustly Optimized BERT Pretraining Approach (RoBERTa) [29]. These methods process all elements simultaneously by forming direct connections between individual elements through a process known as attention. BERT learns multiple attention mechanisms, called heads, which operate in parallel to one another. The model captures a broader range of relationships between the words via multi-head attention. It is trained based on transfer learning because the model is first trained on two unsupervised tasks. The first one is based on Masked Language Modeling (MLM), to predict missing words (masked) in a sentence. The second one is Next Sentence Prediction (NSP), where the system is trained to predict whether a sentence follows another. RoBERTa uses the same concept as BERT, but removes the NSP part and uses larger batch sizes. In addition, this method introduces dynamic masking for the MLM, which masked tokens change during the training epochs. We considered the BERT-Base model trained with the Wikicorpus data from 102 languages and RoBERTA-Base model trained with filtered CommonCrawl data from 100 languages [30]. The last layer (768 units) is taken as the word-embedding representation in both methods. The mean of the overall word-embeddings is computed for the classification task. The source code is also available online1 [31].

2.2.2. Acoustic Embeddings

These embeddings are created with the Wav2Vec 2.0 [32] model, which aims to learn representations from raw speech signals using self-supervised methods. Wav2Vec’s initial idea was to take raw audio as input and then compute an encoded representation to be used as input for a speech recognition system. The model consists of three main parts: feature extraction, a context network, and a linear projection to the output. The input consists of 16 kHz raw audio divided into chunks of 25 ms. The feature extraction part is composed of several temporal convolutions which convert the speech input into a latent space representation. Similar to the masking idea used in BERT for the self-supervised training, segments of audio are masked (instead of words), and then quantized. Further, a contextualized representation is obtained through a combination of a “Transformers”-based approach with contrastive learning. This approach

---

1https://github.com/PauPerezT/WEBERT
has been fine-tuned in a broader range of tasks, covering from emotions recognition and speaker verification to the assessment of speech disorders. We considered two different pre-trained Wav2Vec models: XLSR-53 [33], which is a multilingual approach trained with CommonVoice, BABEL, and Multilingual LibriSpeech (MLS) datasets; and Wav2Vec for Xvectors [34], which consists of an XVector feature extraction head on top, which was trained with the VoxCeleb corpus [35]. For the XLSR-53 model, we used the output of the last convolutional layer of the model (number of chunks × 512) and computed the mean value over the chunks (512 dimensions). We do not consider the embeddings extracted from the last layer because they did not provide relevant results to this approach. The resulting embedding obtained from the Wav2Vec for Xvectors was fixed per speaker and had a total of 512 dimensions.

3. Experiments and Results

Two different experiments were performed in this work: (1) the classification of AD vs. HC using the pre-trained embeddings and (2) transfer learning using English data to enhance the AD discrimination in Spanish data. Three different classifiers were considered for comparison purposes: a Radial basis function Support Vector Machine (SVM), XGBoost, and a Multi Layer Perceptron (MLP). The classifiers were optimized following a 4-fold cross-validation strategy. The optimal parameters of each classifier were found through a grid search, where for the SVM the number of estimators \( \gamma \) and the maximum depth \( \gamma \) and the maximum depth \( \gamma \in \{ 10^{-4}, 10^{-3}, ..., 1 \} \), for the XGBoost the number of estimators \( \gamma \in \{ 50, 70, ..., 300 \} \) and the maximum depth \( \gamma \in \{ 1, 3, 5, 7 \} \). For the MLP only a single layer of 768 neurons was considered. Since the folds were randomly chosen and label-balanced, the classification was performed 10 times.

3.1. Classification using the pre-trained embeddings

We performed a visual inspection by using Principal Component Analysis (PCA). We noticed strong language dependencies, even though the embeddings were trained with multilingual data. Figure 1 shows an example for the case of the XLSR-53 model.

![Figure 1: Principal component analysis projection of the original XLSR-53 Wav2Vec embeddings for each target group. ES: Spanish. EN: English](image)

Classification results are shown in Table 3, where each embedding was considered separately, and a comparison between different classifiers and languages was performed. The classification is performed considering English only, Spanish only, and the combination of English and Spanish. Since there are more samples for English than for Spanish, the number of samples was balanced for training and testing.

<table>
<thead>
<tr>
<th>Source</th>
<th>Emb Language</th>
<th>SVM</th>
<th>XGBoost</th>
<th>MLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistic</td>
<td>BT</td>
<td>EN</td>
<td>0.68 (0.01)</td>
<td>0.69 (0.02)</td>
</tr>
<tr>
<td></td>
<td>ES</td>
<td>0.53 (0.08)</td>
<td>0.46 (0.14)</td>
<td>0.51 (0.10)</td>
</tr>
<tr>
<td></td>
<td>EN-ES</td>
<td>0.60 (0.06)</td>
<td>0.54 (0.02)</td>
<td>0.52 (0.03)</td>
</tr>
<tr>
<td></td>
<td>EN</td>
<td>0.63 (0.03)</td>
<td>0.74 (0.01)</td>
<td>0.76 (0.02)</td>
</tr>
<tr>
<td></td>
<td>RFT</td>
<td>ES</td>
<td>0.61 (0.06)</td>
<td>0.52 (0.20)</td>
</tr>
<tr>
<td></td>
<td>EN-ES</td>
<td>0.62 (0.07)</td>
<td>0.54 (0.07)</td>
<td>0.64 (0.06)</td>
</tr>
<tr>
<td>Acoustic</td>
<td>W2</td>
<td>EN</td>
<td>0.62 (0.04)</td>
<td>0.53 (0.03)</td>
</tr>
<tr>
<td></td>
<td>ES</td>
<td>0.79 (0.06)</td>
<td>0.78 (0.04)</td>
<td>0.80 (0.05)</td>
</tr>
<tr>
<td></td>
<td>EN-ES</td>
<td>0.67 (0.03)</td>
<td>0.64 (0.04)</td>
<td>0.67 (0.02)</td>
</tr>
<tr>
<td>Acoustic</td>
<td>WX</td>
<td>EN</td>
<td>0.63 (0.04)</td>
<td>0.55 (0.01)</td>
</tr>
<tr>
<td></td>
<td>ES</td>
<td>0.58 (0.07)</td>
<td>0.69 (0.05)</td>
<td>0.70 (0.07)</td>
</tr>
<tr>
<td></td>
<td>EN-ES</td>
<td>0.58 (0.10)</td>
<td>0.56 (0.03)</td>
<td>0.60 (0.05)</td>
</tr>
</tbody>
</table>


Notice that the highest performance for discriminating AD in English speakers was obtained by using linguistic embeddings (RoBERTa = 0.76), while for Spanish were the acoustic embeddings (XLSR-53 Wav2Vec = 0.80). Regarding the classifiers, SVM and MLP yielded similar results. The classification of AD vs. HC was also performed combining the two datasets, achieving lower results compared to those obtained with each language separately.

3.2. Transfer knowledge from English to Spanish

A transfer learning based approach was proposed to enhance the classification of AD.

Figure 2 shows the general scheme followed for transferring information from English to Spanish, where:

1. Training of the transfer learning model was performed with the English corpus. It was divided into 80% training, 10% validation, and 10% testing.
2. The encoded representation of the signal/text is computed using the pre-trained models.
3. Each pre-trained embedding was fine-tuned to discriminate AD in English speakers. We used a linear layer with the same dimensionality as the input (512 for acoustic, 768 for linguistic) along with a GELU activation, and a classification layer.
4. For “projecting” the pre-trained embedding onto the Spanish data, we took the output of the linear layer (trained in English data) before the classification layer. Later on, these new embeddings were used for classifying AD in the Spanish data.

The main assumption is that adjusting the embeddings to a different task, despite being in a different language, improves the performance. By visual inspection using PCA in Figure 3, it can be observed that the projected embeddings improved the separation between samples while maintaining good inter-class separability. This type of behavior can be useful for the classifier, which can better distinguish classes using the projected space.

Table 4 shows the results using the projected embedding obtained from the transfer learning for classifying AD in Spanish speakers. A slight improvement is observed for the XLSR-53 model; however, the rest of the results show a bigger improvement for AD discrimination. Although the performance increased for the linguistic models, the acoustic model is still the one with more accurate results.

Table 5: Best classification results obtained with the combination of the different pre-trained embeddings. EN: Pitt corpus. ES: Chilean Spanish corpus. EN + ES: Pitt and Chilean Spanish corpora together. EN → ES: transfer learning from the Pitt to the Chilean Spanish corpus (Spanish only)

<table>
<thead>
<tr>
<th>Language</th>
<th>Emb</th>
<th>Classifier</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>RBT+W2V</td>
<td>MLP</td>
<td>0.81 (0.02)</td>
</tr>
<tr>
<td>ES</td>
<td>W2+WX</td>
<td>SVM</td>
<td>0.81 (0.03)</td>
</tr>
<tr>
<td>EN + ES</td>
<td>RBT+W2</td>
<td>MLP</td>
<td>0.78 (0.01)</td>
</tr>
<tr>
<td>EN → ES</td>
<td>RBT+W2+WX</td>
<td>SVM</td>
<td>0.85 (0.01)</td>
</tr>
</tbody>
</table>

Values are expressed as mean (standard deviation). Emb: Embeddings.
WX: Wav2Vec for Xvectors model.

Moreover, the fusion of embeddings helped in improving the classification for all the language combinations.

4. Discussion and Conclusions

The use of multilingual embeddings along with a transfer learning-based approach was proposed for a cross-lingual analysis of AD. The aim was to improve inter-class separability while exploring the possibility of using these embeddings in multilingual analyses. As expected, strong language dependencies were observed, especially for linguistics embeddings, which also performed better in English. A lower language dependency was observed for the acoustics embeddings, but unlike linguistics, it achieved better results on the Spanish corpus. We believe that this is due to the fact that the phonetic “space” of sounds in Spanish is simpler and more predictable than in English. Spanish contains 25 phonemes while English has 44 phonemes [22] and the structure of the first is mainly based on the consonant-vowel combination, while the second includes also several consonant clusters. Besides, we want to highlight that Wav2Vec results should be taken with caution, as they may be markedly influenced by acquisition conditions. Further research is required in these two aspects hopefully with more data to validate our observations. Another contribution of this work is the use of a transfer learning approach to improve the separability between AD and HC classes. Our numerical and visual findings suggest that by transferring information of the same task from English to Spanish improves the inter-classes separability. The results with the linguistic models suggest that more exploration is needed due to the grammatical and semantic complexity of each language. We are aware of the limitations of this study, mainly regarding the need for additional data and also for a systematic evaluation of possible acoustic differences among recording sessions. Future work will consider different features that may be useful for cross-linguistic analyses in AD, especially those that show higher inter-class separability and less inter-language dependency. Other fusion methods will be explored to see whether classification accuracies improve.

5. Acknowledgements

This study was partially funded by CODI at UdeA grant # PRG2020-34068. Adolfo García received funding from the Global Brain Health Institute (GBHI), Alzheimer’s Association, and Alzheimer’s Society (Alzheimer’s Association GBHI ALZ UK-22-865742); ANID, FONDECYT Regular [1210176]; and Programa Interdisciplinario de Investigación Experimental en Comunicación y Cognición (PIIECC), Facultad de Humanidades, USACH.
6. References


