An Automatic System for Dementia Detection using Acoustic and Linguistic Features

Miriam Gonzalez-Atienza, Jose A. Gonzalez-Lopez, and Antonio M. Peinado

Dept. of Signal Theory, Telematics and Communications
University of Granada
myriamgonzalez@correo.ugr.es, joseangl@ugr.es, amp@ugr.es

Abstract

Early diagnosis of dementia is crucial for mitigating the consequences of this disease in patients. Previous studies have demonstrated that it is possible to detect the symptoms of dementia, in some cases even years before the onset of the disease, by detecting neurodegeneration-associated characteristics in a person’s speech. This paper presents an automatic method for detecting dementia caused by Alzheimer’s disease (AD) through a wide range of acoustic and linguistic features extracted from the person’s speech. Two well-known databases containing speech for patients with AD and healthy controls are used to this end: DementiaBank and ADReSS. The experimental results show that our system is able to achieve state-of-the-art performance on both databases. Furthermore, our results also show that the linguistic features extracted from the speech transcription are significantly better for detecting dementia.

Index Terms: Acoustic voice analysis, speech-based disease diagnosis, dementia, Alzheimer’s disease, word embeddings.

1. Introduction

Dementia is a type of neurodegenerative disease, whose most common cause is Alzheimer’s Disease (AD). Although memory impairment is the main early symptom for AD, it has been found that language and speech abilities also decline, even in the very early stages of the disease [1]. This is known to affect object naming, noun production and rates of verb usage. In general, loss of vocabulary, simplified syntax/semantics, and overuse of semantically empty words are commonly found in the language of people with dementia [2, 3]. Speech is, therefore, a promising candidate as a source of information for new approaches to diagnosing dementia. As no curative therapy is known, early secondary prevention measures are of great importance. Examinations typically include a large number of neuropsychological tests, which are very time consuming and expensive. In order to enable longitudinal cognitive status monitoring on a large scale, a fast and cheap method for diagnosing the disease needs to be found.

Automatic speech processing has been shown to be a promising way in the diagnosis of dementia. Approaches have used acoustic, prosodic and linguistic features [4, 5, 6] in a classification task that aims to distinguish people affected by dementia from cognitively healthy subjects using just their speech. Patient’s speech and language are obtained from written texts and speech recordings. For example, picture description is a constrained task that relies less on episodic memory, but requires more semantic knowledge and retrieval ability. The most commonly used picture prompt is a line drawing called "Cookie Theft" [7]. During the test, the patients are asked to describe what they see in the picture, while the answer is recorded. As reported [8], people suffering from dementia tend to hesitate more often and make longer pauses. Thus, features based on the occurrence and duration of pauses, extracted from the output of a Voice Activity Detector (VAD), are very promising for automatic detection of this disease.

In this paper, we present the details of our system for dementia detection using features extracted from speech recordings and its corresponding transcriptions. We use both acoustic and linguistic features and compare the results to those obtained in previous investigations. Feature selection of the most relevant features is performed in order to achieve better classification results. We evaluated our proposed system on two well-known databases containing speech material recorded by AD patients and healthy controls (HC): DementiaBank [9] and the recently released Alzheimer’s dementia recognition through spontaneous speech (ADReSS) database [10].

This paper has the following structure. In Section 2, we present the details of our system for automatic dementia detection. Section 3 is dedicated to explain the process followed to select the most important features in classification. The experimental results are presented in Section 5, whereas the main conclusions of this work are listed in Section 7.

2. Dementia detection system

Figure 1 shows a block diagram for the proposed automatic system for speech-based dementia detection. Firstly, the voice of the participant is recorded and transcribed (manually, in our case) while the subject performs a cognitive task (i.e., describing a drawing such as the Cookie Theft from the Boston test). Then, a noise reduction technique is applied to the audio signal. The enhanced signal is passed to the feature extraction block where a set of acoustic and linguistic features are extracted. Finally, a selection of these features are sent to a machine learning classifier trained to discriminate between AD and HC subjects.

The details of our system are given in the following sections. Table 1 shows a summary of the acoustic and linguistic features extracted by our system for automatic dementia detection.

2.1. Acoustic features

Acoustic features measure how participants speak. They are extracted from the denoised speech signals after applying the spectral noise gating method described in [11] to the audio signals to improve their quality. In our system, we consider the

---

This work was funded by the Spanish State Research Agency (SRA) under the grant PID2019-108040RB-C22/SRA/10.13039/501100011033. Jose A. Gonzalez-Lopez holds a Juan de la Cierva-Incorporation Fellowship from the Spanish Ministry of Science, Innovation and Universities (IJCI-2017-33926).
following four types of acoustic features:

Pause-based features: People afflicted with dementia tend to hesitate more often and make longer pauses than HC subjects [6, 12]. Thus, we computed four pause-based features for each utterance, including total pause and speech duration (in seconds), rate between total duration of pauses and duration of the utterance, and rate between the total duration of pauses and the duration of speech for each utterance. These features were automatically computed from the outputs of an energy-based VAD technique applied to the de-noised signals. The energy threshold was manually defined as 60% of the signal energy.

Speech rhythm: Longer pauses and more hesitations in speech from people with dementia imply that more time is needed to convey words. Thus, we measured the speech rhythm as number of beats per minute using the Predominant Local Pulse (PLP) algorithm, first introduced by Grosche and Müller [13].

Spectral features: Mel-frequency cepstral coefficients (MFCCs) [14, 11, 15, 16] were extracted from windows of 25 ms with 15 ms overlap to capture the spectral content of the speech signal. We then computed the average across time for the first 24 MFCCs. Then, we computed the mean, variance, skewness and kurtosis of the resulting average vector. We also extracted their first order derivatives with their mean and variance. Spectral centroid parameters [17], which aim to locate the spectrum centre of mass and have been found to be valuable in measuring the cognitive load [18, 11], were also extracted from the audio waveforms and computed their mean.

Prosodic features: We computed the speech fundamental frequency ($F_0$) using the autocorrelation method [19]. From the $F_0$ values, we computed the mean and standard deviation through the signal.

2.2. Linguistic features

Linguistic features are used to measure changes in vocabulary and sentence structure that are caused by dementia. Our linguistic features operate at the word level of transcriptions. In this work, the manual transcriptions provided with the databases are used to extract the linguistic features. In general, AD patients tend to make shorter phrases than HC and also have a less-rich vocabulary. The linguistic features described below are computed as proportions considering the total number of words spoken in each utterance, e.g., percentage of adjectives w.r.t. the total number of spoken words.

Word count: Word count is a common used feature to classify people with dementia [16]. Verbal repetition is a hallmark of dementia and AD at all stages, but is most commonly targeted for monitoring and treatment effects in its mild stage [20]. For that reason, we have extracted the proportion between unique words and the total number of words in each transcription.

Part-of-Speech (POS) Tags: Words with similar grammatical properties can be grouped together by POS tags. Each tag represents the grammatical role a word can take in a sentence and thus POS tags can be used to indicate grammatical properties of participant’s speech. We used TreeTagger [21] to automatically extract POS tags and calculate the frequency of occurrence of each tag. We have considered three POS categories: verbs, determinants and nouns. Once the frequency of occurrence is calculated for each participant, we measure the proportion between the number of POS categories and the number of words in each sentence.

Word embeddings: Word embedding is a technique widely used to convert written words into feature vectors. Recently, successful approaches have used deep learning techniques to produce vectors representing words. In this work, we have used the word2vec technique [22], which is based on the co-occurrences of words taking into account the context, to extract 300-dimensional embeddings for every word spoken by the participant in each sentence. Then, we average all the embeddings obtained in each sentence to finally obtain a 300-dimensional vector representing the linguistic content of that sentence. Besides, we calculate the mean and standard deviation for each word embedding vector of the subject. Thus, we obtain a 300-dimensional vector for each subject, with two additional values: the mean and standard deviation of this vector.

3. Feature selection

We applied a feature selection procedure to select the most meaningful acoustic and linguistic features and discard the features that contribute less to the classification accuracy. We trained an extremely randomized trees (ERT) classifier [23], which is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a “forest” to output its classification result. Feature selection was applied to the training set and then, the optimum
Feature importance

To evaluate the proposed dementia detection system, we used the audio recordings and transcriptions from the DementiaBank [11, 24] and ADReSS [10] databases. DementiaBank is a free-access, large existing database for Alzheimer’s and related dementia diseases collected during longitudinal study conducted by the University of Pittsburgh School of Medicine. Verbal descriptions of the Boston Cookie Theft picture were recorded from people with different types of dementia with an age span from 49 to 90 years as well as from elderly healthy control subjects within an age range from 46 to 81 years. During the interviews, patients were given the picture and were told to discuss everything they could see happening in the picture. There are a total of 473 recordings from 97 healthy controls and 233 speech samples from 167 AD patients diagnosed as possible or probable AD. We split this database randomly into training and evaluation subsets. In the training subset, we have selected 150 subjects, as well as for the evaluation subset, in order to have the same number of subject in both groups.

Similarly, the ADReSS database is a subset of DementiaBank with acoustically enhanced audio recordings and matched in terms of age and gender (i.e., major factors in recognising dementia) to avoid bias towards them. The dataset contains speech recordings from 78 non-AD subjects and 78 AD subjects while describing the Cookie Theft drawing. This database already defines a training subset, containing data for 54 AD and 54 HC subjects, and an evaluation subset with 24 subjects for each group.

A summary of the characteristics of both databases in terms of number of subjects in each class is shown in Table 2.

Table 2: Basic characteristics of the patients in each group in the ADReSS challenge dataset and DementiaBank

<table>
<thead>
<tr>
<th>Dataset</th>
<th>No. subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AD</td>
</tr>
<tr>
<td>ADReSS Train</td>
<td>54</td>
</tr>
<tr>
<td>ADReSS Test</td>
<td>24</td>
</tr>
<tr>
<td>DementiaBank</td>
<td>322</td>
</tr>
</tbody>
</table>

4. Experimental setup

4.1. Databases

To evaluate the proposed dementia detection system, we used the audio recordings and transcriptions from the DementiaBank [11, 24] and ADReSS [10] databases. DementiaBank is a free-access, large existing database for Alzheimer’s and related dementia diseases collected during longitudinal study conducted...
From the comparison between acoustic and linguistic features, we conclude that linguistic features provide significantly better classification results. Hence, they present more capability to distinguish between HC and AD than acoustic features. Recently and similar to our work, Fraser et al. [16] studied the potential of using linguistic features to identify Alzheimer's disease. In total, a set of 370 acoustic, lexical and semantic features were extracted and they obtained a highest accuracy of 92% in distinguishing between HC subjects and AD patients using the top 25 ranked features.

From the comparison between the results obtained on the ADReSS and DementiaBank datasets, it can be observed that better classification results are obtained on DementiaBank. One possible explanation is the number of training subjects in each dataset, which is considerably larger in DementiaBank (300 subject in DementiaBank vs. 108 in ADReSS). Another possible explanation could be that ADReSS is gender and age balanced, whereas DementiaBank is not. Thus, it could be that, in the case of DementiaBank, the classifiers are able to infer the age and gender of each subject from the e.g., acoustic features to achieve a better performance. Besides, the recordings from DementiaBank have a higher level of background noise than the recordings from ADReSS, which causes that acoustic features contribute less to the classification results in comparison with the acoustic features on ADReSS dataset. Finally, as can be seen from Table 3 and 4, we have achieved higher classification results than the ones presented in the literature.

7. Conclusions

In this paper, we have proposed a binary classifier approach based on speech for dementia detection using acoustic and linguistic features derived from audio recordings and text transcriptions. Feature Selection has showed that the best features for this task are pause-based features, word embeddings and Conversational POS categories (nouns, verbs and determinants). We have evaluated our system on two well-known databases achieving state-of-the-art results for both. We have obtained the best classification results with LDA classifier and linguistic features for DementiaBank dataset, achieving an accuracy of 96%, a sensitivity of 97% and 92% of specificity. On the other hand, with ADReSS dataset we have achieved the maximum classification results with SVM classifier and 15% of the features.

Table 3: Classification results for DementiaBank dataset. Accuracy/Sensitivity/Specificity (%). In bold are shown the higher classification results by rows

<table>
<thead>
<tr>
<th>Features</th>
<th>LDA</th>
<th>SVM</th>
<th>RF</th>
<th>AdaBoost</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistic (with word embeddings)</td>
<td>96/97/92</td>
<td>62/62/62</td>
<td>74/74/74</td>
<td>78/78/78</td>
<td>78/78/78</td>
</tr>
<tr>
<td>Feature selection (15%)</td>
<td>95/95/95</td>
<td>67/67/67</td>
<td>78/78/78</td>
<td>81/81/81</td>
<td>80/80/80</td>
</tr>
<tr>
<td>Feature selection (25%)</td>
<td>93/93/93</td>
<td>75/75/75</td>
<td>81/81/81</td>
<td>87/87/87</td>
<td>82/82/82</td>
</tr>
<tr>
<td>Feature selection (50%)</td>
<td>87/87/87</td>
<td>83/83/83</td>
<td>86/86/86</td>
<td>91/91/91</td>
<td>89/89/89</td>
</tr>
<tr>
<td>Feature selection (75%)</td>
<td>94/94/94</td>
<td>72/72/72</td>
<td>81/81/81</td>
<td>85/85/85</td>
<td>83/83/83</td>
</tr>
<tr>
<td>Feature selection (100%)</td>
<td>82/82/82</td>
<td>71/71/71</td>
<td>82/82/82</td>
<td>85/85/85</td>
<td>83/83/83</td>
</tr>
<tr>
<td>Al-Hammed et al [11, 14]</td>
<td>-</td>
<td>86/-/-</td>
<td>96/-/-</td>
<td>86/-/-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 4: Classification results for ADReSS dataset. Accuracy/Sensitivity/Specificity (%). In bold are shown the higher classification results by rows

<table>
<thead>
<tr>
<th>Features</th>
<th>LDA</th>
<th>SVM</th>
<th>RF</th>
<th>AdaBoost</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acoustic</td>
<td>65/65/65</td>
<td>65/65/65</td>
<td>59/59/59</td>
<td>59/59/59</td>
<td>62/62/62</td>
</tr>
<tr>
<td>Linguistic (with word embeddings)</td>
<td>72/72/72</td>
<td>78/78/78</td>
<td>61/61/61</td>
<td>65/65/65</td>
<td>69/69/69</td>
</tr>
<tr>
<td>Feature selection (15%)</td>
<td>70/70/70</td>
<td>79/79/79</td>
<td>57/57/57</td>
<td>57/57/57</td>
<td>61/61/61</td>
</tr>
<tr>
<td>Feature selection (25%)</td>
<td>53/53/53</td>
<td>66/66/66</td>
<td>57/57/57</td>
<td>57/57/57</td>
<td>58/58/58</td>
</tr>
<tr>
<td>Feature selection (50%)</td>
<td>59/59/59</td>
<td>55/55/55</td>
<td>69/69/69</td>
<td>57/57/57</td>
<td>60/60/60</td>
</tr>
<tr>
<td>Feature selection (75%)</td>
<td>65/65/65</td>
<td>61/61/61</td>
<td>58/58/58</td>
<td>56/56/56</td>
<td>60/60/60</td>
</tr>
<tr>
<td>Feature selection (100%)</td>
<td>60/60/60</td>
<td>62/62/62</td>
<td>66/66/66</td>
<td>57/57/57</td>
<td>62/62/62</td>
</tr>
<tr>
<td>Martinc et al. [28]</td>
<td>77/-/-</td>
<td>51/-/-</td>
<td>55/-/-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Luz et al. [10]</td>
<td>-</td>
<td>75/-/-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

6. Discussion

From the comparison between acoustic and linguistic features, we conclude that linguistic features provide significantly better classification results. Consequently, they present more capability to distinguish between HC and AD than acoustic features. Recently and similar to our work, Fraser et al. [16] studied the potential of using linguistic features to identify Alzheimer's disease. In total, a set of 370 acoustic, lexical and semantic features were extracted and they obtained a highest accuracy of 92% in distinguishing between HC subjects and AD patients using the top 25 ranked features.

In the future, we plan to investigate the use of more features as well as other machine learning algorithms. Furthermore, it would be interesting to evaluate the robustness of the system for detecting dementia in other databases or, even, in different languages.
8. References


