ANN-based Alzheimer's disease classification from bag of words

Philipp Klumpp, Julian Fritsch, Elmar Nöth

Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
Email: {philipp.klumpp,julian.fritsch,elmar.noeth}@fau.de

Abstract

Alzheimer's disease (AD) is the most frequent cause of dementia and the patient numbers are increasing within an aging society. Prior research has shown that AD significantly affects the speech signal, and many approaches were published on how to detect AD from only speech or spoken text information. In an earlier work, we have proven the reliability of language models to statistically evaluate transcriptions from AD and healthy control participants. Based on these results, we propose the approach of counting word occurrences in transcriptions, storing them in a bag of words (BoW) vector, and using this vector as an input into an artificial neural network which classifies between AD and healthy state. It could be shown that the new method reached very similar results compared to the language model classifiers, although information about the word order was omitted.

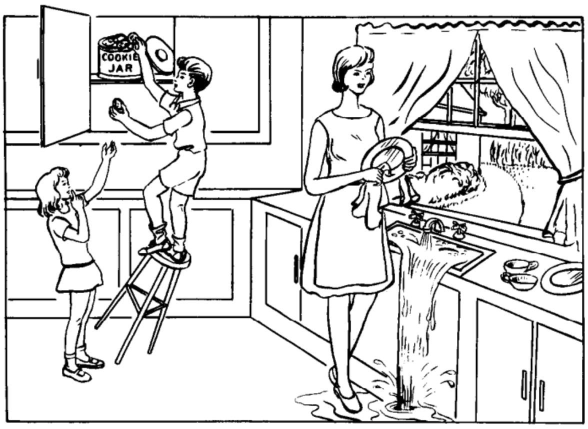

Figure 1: *Cookie Theft* image used for descriptive tasks.

1 Introduction

According to the World Alzheimer Report from 2016, approximately 47 million people are living with dementia worldwide [1]. This number is predicted to increase significantly in the future, mainly as a result of aging societies [2]. Depending on the region, 50 % (Asia) to 70 % (North America) of all dementia cases are diagnosed with Alzheimer's Disease [3]. AD is a neurological disease that mostly affects older people and is characterized by a progressive dementia [4]. It causes a degeneration of specific nerve cells and the presence of neuritic plaques [5].

AD is not to be mistaken with senile dementia, which is commonly developed throughout the aging process [6]. This makes the diagnosis more challenging in early stages of AD. Observing only the symptoms, it can be difficult to differentiate between senile dementia and an early AD. In this case, it might be helpful to provide a tool for automated AD classification which could reduce the workload of a medical expert and provide cheap and highly accessible monitoring.

The majority of research contributions in the field of automated AD classification rely on clinical imaging techniques [7–10]. Nevertheless, it could already be shown that AD affects the speech of patients as well [11]. Several studies proved the reliability of lexical analysis of spontaneous speech for AD [12, 13]. In 1988, Cummings et al. compared the effects of different neurological diseases (Parkinson's Disease and AD) on the human speech signal. They could show that both diseases strongly affect the capabilities of a patient to produce speech, but are still well separable between each other [14]. At Interspeech 2017, Wankerl et al. introduced a statistical approach to classify AD [15]. They computed two n-gram based language models (LM), one for the AD group, another one of the healthy controls. Afterwards, they evaluated the perplexity between a test sample and both language models. This metric defines how well a statistical model predicts a given sample. The reliability of this approach could be further improved in the scope of a Master's Thesis in 2018 by replacing the n-gram language models with long short-term memory (LSTM) recurrent neural network language models [16].

In this work, we evaluate the reliability of AD classification with another deep learning technique. Instead of focusing on local information (a sequence of a few words), we completely discarded this information and focused only on global word occurrences. We were able to show that our method achieved results similar to those of the previously presented LSTM language model approach.

2 Materials and Methods

2.1 Data Set

For all experiments conducted throughout this study, we made use of the Pitt Corpus [17], a data set collected with 194 patients suffering from some form of dementia and 98 healthy controls. The dementia group was reduced to 168 patients, those participants had either been diagnosed with AD or probable AD. The Pitt Corpus provides recordings and transcriptions of numerous exercises. For both the language model as well as the bag of words classification, we made use of the transcriptions of a descriptive task. Participants had been asked to describe the *Cookie Theft* image (Figure 1), a popular speech exercise for neurological diseases [18].

The 98 healthy controls were composed of 67 female and 31 male speakers, who in total contributed 244 transcriptions. 255 transcriptions were available from the AD group of 113 females and 55 males. All of the transcriptions were in English language provided by native English speakers. Before performing the descriptive task, each patient had been assigned a score according to the mini mental state examination (MMSE), a measure for the cognitive abilities of a person [19]. Several participants contributed more than only one transcription over a longer period of time, therefore they had their MMSE scores measured mul-

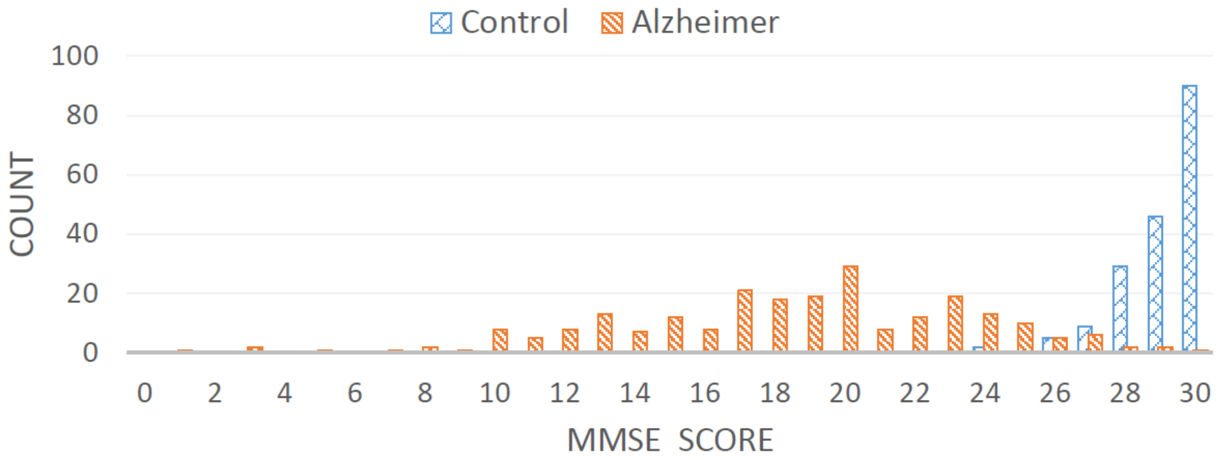

Figure 2: Occurrences of MMSE scores over all transcriptions in the Pitt Corpus.

tiple times. The distribution of MMSE scores for the transcriptions of both the healthy control group as well as the AD patients is shown in Figure 2.

2.2 Language Model Classification

In this section, we want to describe briefly how the language model classification was realized. For a more detailed insight, we refer to related publications [15, 16, 20]. Language models provide a statistical representation for word combinations, and they are language dependent. For the LM classification, two independent language models were computed from the transcriptions. The assumption behind the LM approach was that a cognitive impairment caused by AD would have influence on the ability of a patient to produce meaningful sentences. The intelligibility of each word could remain unaffected, unlike for Parkinson's Disease for example [21], but the production of meaningful word sequences might suffer with dementia. From the transcriptions with simulated 100 % word recognition, two LMs were computed. Both were implemented as recurrent neural networks with LSTM cells. The first LM represented all the transcriptions of the control group, the second one was computed for the AD patients' image descriptions. To classify a new sample, the perplexity of the sample was computed on both LMs. A lower perplexity indicated that the corresponding LM would better predict the sample. The difference between the perplexities with the AD and the healthy control language model was computed and afterwards compared to a threshold determined during training.

2.3 Bag of Words Classification

This work focused on the question whether it is possible to omit the information about the general structure of language in the transcriptions, and still classify AD with a satisfying reliability. A bag of words vector was created which contained 546 values. Each value represented the absolute number of occurrences of one out of 546 stemmed words using the Stanford CoreNLP Toolkit [22]. Stemming the words was important, because we intended to count occurrences of basic word forms, and not distinguish between singular and plural for example. As the vector was composed of all possible words from the 499 transliterations, the model was limited to this particular set of words.

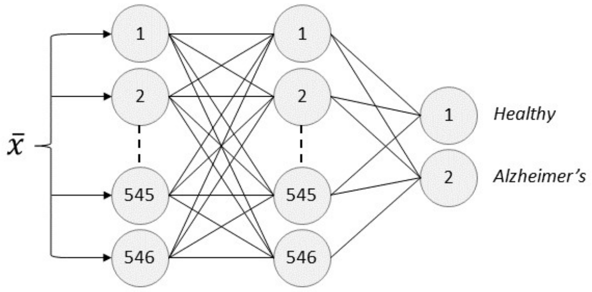

Figure 3: Architecture of the artificial neural network.

The feature vector contained no more information about word sequences. It served as the input for an artificial neural network (ANN). The ANN was realized as a fully connected feed-forward network. It was composed of an input layer with 546 nodes, followed by a hidden layer with another 546 nodes, and finally an output layer with 2 nodes, one for each target class. We applied a softmax function to the output layer, so the probabilities of both classes would always sum up to 1.

The network was trained with a batch size of 20 samples that were randomly chosen from the training set. To prevent the ANN from overfitting, we applied a dropout between the input layer and the hidden layer, throwing away 80 % of the input values. All nodes in the network incorporated a Rectified Linear Unit (ReLU) activation function.

2.4 Evaluation

Both classifiers were evaluated in a leave-one-speaker-out cross-validation. Leaving one speaker out instead of only one sample is important, otherwise the networks would learn knowledge about a particular speaker from the other samples of the candidate. This would have violated the requirement of generalization. We classified each sample of every speaker after training the network with the remaining speakers and afterwards evaluated the overall accuracy of the ANN. Furthermore, we determined a confusion matrix from the classification results of the network.

For both approaches, we computed a receiver operating characteristics (ROC) curve to compare the performance of the classifiers. Additionally, the area under the ROC

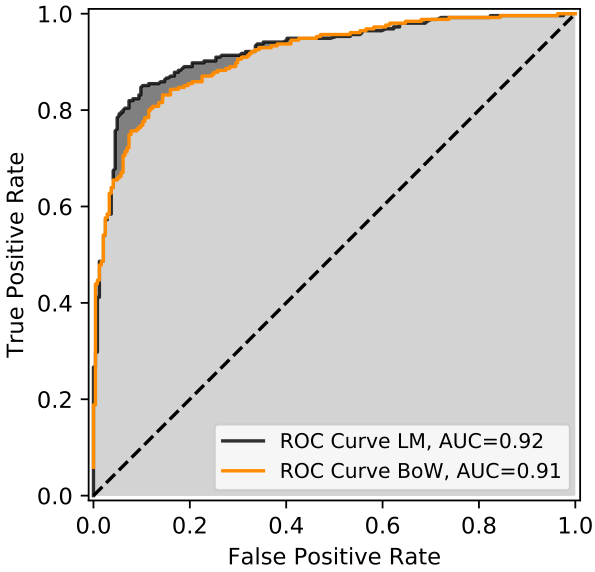

Figure 4: ROC curves of the Bag of Words and Language Model approaches.

curve (AUC) served as a measure to evaluate the methods. To evaluate the relationship between MMSE scores and predicted probability of AD, the Pearson and Spearman correlation coefficients were calculated on the whole data set as well as for the AD group only.

For the BoW method, we also visualized the relative amount of misclassified samples depending on their MMSE score. Finally, we also analyzed the sets of wrongly classified transcriptions of both methods to determine the amount of overlap between the results. For this analysis, we used thresholds for each method to classify at equal error rate.

3 Results

Figure 4 shows the ROC curves of both the LM as well as the BoW approach. Both approaches yield significantly better results compared to random guessing (dashed line). The AUC values were 0.92 for the LM and 0.91 for the BoW approach. Between 0.05 and 0.25 on the x-axis, the LM classifier slightly outperforms the results of the BoW classification. The overall classification accuracy of the BoW method was 84.4 %.

In Figure 5, the relative amount of classification errors were plotted for every MMSE score. It can be seen that with an increasing MMSE score, the participants of the AD group we classified less accurately. For the healthy control group, the amount of misclassified samples increases as the MMSE score is decreasing.

The correlation coefficients for each method are listed in Table 1. In general, the LM approach results correlate better with the MMSE scores of the participants. This difference becomes more clear when only looking at the results of the Alzheimer group. Unlike the LM results, the BoW probabilities were negatively correlated to the MMSE score. An increase in AD probability correlated to a decrease in MMSE.

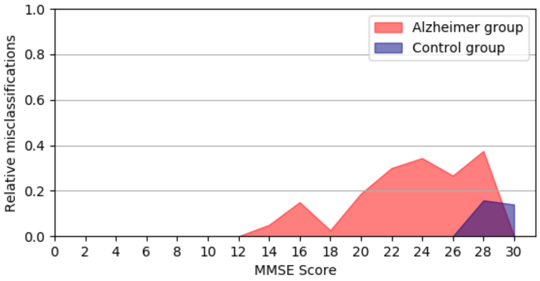

Figure 5: Relative amount of misclassified samples of AD patients and healthy controls sorted by MMSE score.

Table 1: The Pearson and Spearman correlation coefficients computed for both approaches

	Correlation Pearson	Correlation Spearman	
Alzheimer	-0.33	-0.44	BoW
All	-0.68	-0.73	
Alzheimer	0.43	0.55	LM
All	0.66	0.77	

In total, the LM method produced 36 false negative and 36 false positive misclassifications. The BoW method produced 40 false negative and 39 false positive results at equal error rate. The intersecting set of the false negatives was larger with an overlap of 22 transcriptions, compared to the 14 overlapping transcriptions from the false positives.

4 Discussion

Comparing the two ROC curves of the different approaches, it can be seen that the overall results are very similar to each other. The LM method is still slightly better in some regions, but the BoW network comes very close. This similarity is also shown by the AUC values of both approaches, which only differ by 0.01.

As it was shown in Figure 5, the relative amount of classification errors strongly depended on the MMSE score of the observed transcriptions. Participants who had been diagnosed with AD in an early stage reached a relatively high MMSE score. Their mental capabilities were not yet as affected as it could be observed for patients with lower MMSE. Therefore, it was hard for the system to classify them correctly. This problem becomes even more pronounced when looking at the group of healthy controls. None of them had been diagnosed with AD, but according to Table 2, some of them did not achieve an MMSE of 30 anymore. For example, this could be caused by senile dementia. From Figure 5 it can be seen that the relative amount of misclassifications decreased noticeably as the MMSE score of the participants decreased.

The correlations of both methods with the MMSE scores show that the LM classifier worked better for predicting the individual state of a patient. This difference was particularly pronounced when we investigated the correlation for the subgroup of AD participants only. The difference can be explained by the setup of the two methods. For the LM classifier, the difference between two perplexities had been computed. Therefore, the value contained information about which language model would better represent a transcription. For the BoW on the other side, the ANN computed only a probability of a transcription to belong to one class or another. The network never learned about any MMSE scores, it solely learned to perform a binary decision between healthy and AD.

The size of the intersecting sets indicated that there are numerous transcriptions which were misclassified by both methods. However, for a majority of both false positive and false negative classifications, at least one of the classifiers yielded a correct result. This was expected to a certain extent, because the LM approach is reliable for analyzing local language properties, whilst the BoW method on the other hand performed a global language analysis.

References

[1] M. Prince, A. Comas-Herrera, M. Knapp, M. Guerchet, and M. Karagiannidou, "World alzheimer report 2016: improving healthcare for people living with dementia: coverage, quality and costs now and in the future," 2016.

[2] A. Wimo, B. Winblad, H. Aguero-Torres, and E. von Strauss, "The magnitude of dementia occurrence in the world," *Alzheimer Disease & Associated Disorders*, vol. 17, no. 2, pp. 63–67, 2003.

[3] L. Fratiglioni, D. De Ronchi, and H. Agüero-Torres, "Worldwide prevalence and incidence of dementia," *Drugs & aging*, vol. 15, no. 5, pp. 365–375, 1999.

[4] C. P. Ferri, M. Prince, C. Brayne, H. Brodaty, L. Fratiglioni, M. Ganguli, K. Hall, K. Hasegawa, H. Hendrie, Y. Huang, *et al.*, "Global prevalence of dementia: a delphi consensus study," *The lancet*, vol. 366, no. 9503, pp. 2112–2117, 2005.

[5] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E. M. Stadlan, "Clinical diagnosis of alzheimer's disease report of the nincds-adrda work group* under the auspices of department of health and human services task force on alzheimer's disease," *Neurology*, vol. 34, no. 7, pp. 939–939, 1984.

[6] C. Gottfries, "Alzheimer's disease and senile dementia: biochemical characteristics and aspects of treatment," *Psychopharmacology*, vol. 86, no. 3, pp. 245–252, 1985.

[7] L. O'Dwyer, F. Lamberton, A. L. Bokde, M. Ewers, Y. O. Faluyi, C. Tanner, B. Mazoyer, D. O'Neill, M. Bartley, D. R. Collins, *et al.*, "Using support vector machines with multiple indices of diffusion for automated classification of mild cognitive impairment," *PloS one*, vol. 7, no. 2, p. e32441, 2012.

[8] E. Westman, J.-S. Muehlboeck, and A. Simmons, "Combining mri and csf measures for classification of alzheimer's disease and prediction of mild cognitive impairment conversion," *Neuroimage*, vol. 62, no. 1, pp. 229–238, 2012.

[9] Y. Zhang, S. Wang, and Z. Dong, "Classification of alzheimer disease based on structural magnetic resonance imaging by kernel support vector machine decision tree," *Progress In Electromagnetics Research*, vol. 144, pp. 171–184, 2014.

[10] S. Klöppel, C. M. Stonnington, C. Chu, B. Draganski, R. I. Scahill, J. D. Rohrer, N. C. Fox, C. R. Jack Jr, J. Ashburner, and R. S. Frackowiak, "Automatic classification of mr scans in alzheimer's disease," *Brain*, vol. 131, no. 3, pp. 681–689, 2008.

[11] M. Nicholas, L. K. Obler, M. L. Albert, and N. Helm-Estabrooks, "Empty speech in alzheimer's disease and fluent aphasia," *Journal of Speech, Language, and Hearing Research*, vol. 28, no. 3, pp. 405–410, 1985.

[12] R. S. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock, "Analysis of spontaneous, conversational speech in dementia of alzheimer type: Evaluation of an objective technique for analysing lexical performance," *Aphasiology*, vol. 14, no. 1, pp. 71–91, 2000.

[13] C. Thomas, V. Keselj, N. Cercone, K. Rockwood, and E. Asp, "Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech," in *Mechatronics and Automation, 2005 IEEE International Conference*, vol. 3, pp. 1569–1574, IEEE, 2005.

[14] J. L. Cummings, A. Darkins, M. Mendez, M. A. Hill, and D. Benson, "Alzheimer's disease and parkinson's disease comparison of speech and language alterations," *Neurology*, vol. 38, no. 5, pp. 680–680, 1988.

[15] S. Wankerl, E. Nöth, and S. Evert, "An n-gram based approach to the automatic diagnosis of alzheimer's disease from spoken language," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017.

[16] J. Fritsch, "Language Models for Dementia Detection," master's thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg, 2018.

[17] 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.

[18] P. McNamara, L. K. Obler, R. Au, R. Durso, and M. L. Albert, "Speech monitoring skills in alzheimer's disease, parkinson's disease, and normal aging," *Brain and Language*, vol. 42, no. 1, pp. 38–51, 1992.

[19] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""mini-mental state": a practical method for grading the cognitive state of patients for the clinician," *Journal of psychiatric research*, vol. 12, no. 3, pp. 189–198, 1975.

[20] J. Fritsch, C. Bergler, S. Wankerl, and E. Nöth, "Automatic Diagnosis of Alzheimer's Disease Using Neural Networks Language Models," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, p. "to appear", ISCA, 2018.

[21] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with parkinson's disease," *Behavioural neurology*, vol. 11, no. 3, pp. 131–137, 1999.

[22] J. Haiman, *Iconicity in syntax: proceedings of a Symposium on iconicity in syntax, Stanford, June 24-6, 1983*, vol. 6. John Benjamins Publishing, 1985.