INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

An N-Gram Based Approach to the Automatic Diagnosis of Alzheimer's Disease from Spoken Language

Sebastian Wankerl, Elmar Nöth, Stefan Evert

Friedrich-Alexander-University Erlangen-Nuremberg, Germany
{sebastian.wankerl, elmar.noeth, stefan.evert}@fau.de

Abstract

Alzheimer's disease (AD) is the most common cause of dementia and affects wide parts of the elderly population. Since there exists no cure for this illness, it is of particular interest to develop reliable and easy-to-use diagnostic methods to alleviate its effects. Speech can be a useful indicator to reach this goal. We propose a purely statistical approach towards the automatic diagnosis of AD which is solely based on n-gram models with subsequent evaluation of the perplexity and does not incorporate any further linguistic features. Hence, it works independently of a concrete language. We evaluate our approach on the DementiaBank which contains spontaneous speech of test subjects describing a picture. Using the Equal-Error-Rate as classification threshold, we achieve an accuracy of 77.1%. In addition to that, we studied the correlation between the calculated perplexities and the Mini–Mental State Examination (MMSE) scores of the test subjects. While there is little correlation for the healthy control group, a higher correlation could be found when considering the demented speakers. This makes it reasonable to conclude that our approach reveals some of the cognitive limitations of AD patients and can help to better diagnose the disease based on speech.

Index Terms: Pathological speech and language, Applications in medical practice, Automatic analysis of speaker states and traits

1. Introduction

Dementia names the clinical symptom of advancing impairment of memory accompanied by the deterioration of further cognitive functions like the ability to recognize objects or the ability to understand and produce coherent language[1]. It primarily affects senior citizens above the age of 60 where the prevalence ranges between 5%–7% in most parts of the world and within this group, the prevalence increases exponentially with age[2].

The symptom can be caused by various diseases of which Alzheimer's disease (AD) is the most common occurring, being responsible for around two-thirds of dementia cases[3, 4]. AD is a neurodegenerative disease for which no cure is known to the present day. Patients affected by AD, as well as other forms of dementia, usually become dependent on intensive care. Consequently, the various forms of dementia burden the budget of health care systems[5] and drastically reduce the quality of life of the affected persons and their relatives.

To allow an early intervention and best possible treatment for dementia patients, it is of particular interest to develop reliable and easy-to-use diagnostic methods. Currently, most methods for the diagnosis of AD come from the medical domain, like imaging techniques[6], or the psychological domain, e.g. the Mini–Mental State Examination (MMSE)[7]. Since patients of AD commonly show symptoms of anxiety, delusion, or tension[8], it is likely to assume that a medical examination as well as the exam-like atmosphere of the MMSE are sensed as very uncomfortable by the patients. In addition to that, these testing methods are time consuming and expensive.

1.1. Speech and AD

Speech as a daily function can help to provide insights into the condition of a person, too. In particular, it is a source that combines acoustic features with cognitive features like syntactic complexity and vocabulary richness and thereby yields information about both a patient's motoric as well as cognitive capabilities. Moreover, it has the advantage that it is easy to obtain, for instance in a short conversation about a patient's biography or a picture description task.

Previous research suggests language changes in AD patients. Typical patterns include difficulties in naming tasks. This includes difficulties in listing objects belonging to the same category (like animals or furniture)[9, 10] or substituting specific nouns and verbs with more generic or more familiar ones[9, 11] (e.g. *hammer* for *anchor*[9]). Often, words that cannot be remembered at all are substituted with pronouns like *something* which makes the speech of AD patients appearing empty[12]. In addition to that, previously uttered ideas are often repeated[12]. A further characteristic is that AD patients have difficulties with maintaining a conversation, making their speech appear incoherent[13, 14].

1.2. Related Work

Various research has been carried out to computationally diagnose AD from a patient's speech. Weiner and Schultz[15], analyzed a total of 112 hours of conversational speech from 23 test subjects. Each of the 23 test subjects contributed at least two recordings to the corpus and 16 have undergone a change with regard to their cognitive health between two sessions. In total, the used corpus consisted of 51 samples. Considering only acoustic features, the goal was to find possible intra-personal changes in test subjects' health, that the binary decision change↔no change. Using an LDA classifier trained with a cross-validation approach, 80.4% of the samples were classified correctly. More precisely, only few healthy subjects were wrongly attested a change while half of the of the patients who developed a disease were identified correctly.

The study, however, has the drawback that all conclusions were drawn from a relatively small sample size of only 23 participants. Moreover, a quite large amount of recorded speech was used, making it questionable whether this approach also works on shorter samples.

A different approach has been taken by Fraser et. al.[16] who used picture descriptions from the DementiaBank (cf. sec. 3) to distinguish between demented (AD) and healthy elderly. They evaluated a total of 370 features, including features based on syntactic complexity, part-of-speech tagging, vocabu-

http://dx.doi.org/10.21437/Interspeech.2017-1572

lary richness, and acoustics. Using logistic regression and 10-fold cross-validation, the most discriminative features were assessed. These are the features with the highest Pearson correlatin coefficient between features itself and the binary class.

The highest accuracy, 81.92% was obtained when selecting a subset of the 35 highest ranked features. With a subset of 50 features an accuracy of 78.72% was achieved and using all features, it dropped to 58.51%. The drawback of this approach are the large number of features that need to be evaluated as well as the dependence on further language specific tools like part-of-speech taggers or parsers.

2. Theoretical Background

2.1. N-Gram Language Models

N-Gram Language models are a frequently used technique in the processing of both spoken and written language[17, 18]. They generate a probability density from a training text by calculating the frequencies of word sequences. In the simplest case, these sequences consist of only a single word. Here, one counts the words in the training corpus and assigns them an adequate probability. This is called a unigram model. Subsequently, given a sequence $S = (w_1, \dots, w_k)$ of words in a test corpus, one wants to estimate its probability based on the training corpus. In the case of a unigram model, the probability of S, $p(S)$, simply equals the product of the probabilities of each of its words w_i i.e. $p(S) = \prod_{i=1}^{k} p(w_i)$.

However, the drawback of a unigram model is that it doesn't comprise any contextual information, i.e. no information about which words frequently co-occur. On the other hand, considering the whole history of predecessors of each word leads to very unique probabilities. It is intuitively understandable that a sequence of words in the training data which is exceeding a certain length is unlikely to reoccur in the test data as long as both training and test data show a normal variety in their language and are not from a very special domain that only uses limited syntactic constructions and limited vocabulary. Consequently, the approach of remembering the whole history of every word is not practicable.

Taking into account the Markov assumption one can limit the history of each word to a few predecessors. Such a model that contains a history of $n-1$ words is called an n-gram. Moreover, an n-gram that uses a history of length one is called a bigram, for history of length two it is called a trigram. In general, for a sequence $S = (w_1, \dots, w_k)$ of k words and an intended history of length $n-1$, the probability is evaluated by

$$p(S) = \prod_{i=1}^{k} P(w_i | w_{i-n+1}, \dots, w_{i-1}) \tag{1}$$

To calculate the probability of the first words of a sentence, an artificial token which marks the beginning of a sentence is created. The end of a sentence is expressed by a special token, too. It is easy to see that the n-grams incorporate both lexical and syntactic information since the ordering of words is preserved.

When training a language model, one has to keep in mind that the training data is of limited size and most possible n-grams will not be observed. In general, when considering a vocabulary of size N, there are N^2 possible bigrams and N^3 possible trigrams. Thus, for the rather small vocabulary size of 1000, there are already 1,000,000 possible bigrams and 1,000,000,000 possible trigrams. Although this formula is not completely accurate since it factors in syntactically impossible n-grams, too, it is still a useful approximation of the corpus size necessary for observing all n-grams. On the other hand, when considering spontaneous language it is also possible that such syntactically invalid n-grams appear since spontaneous speech is often ungrammatical.

Consequently, it is crucial to smooth, i.e. reshape, the probabilities estimated from relative n-gram frequencies such that n-grams which were not observed in the training data receive a probability greater than zero. Various smoothing techniques are used in literature. In the easiest case, a constant value is added to the frequency of every n-gram (add-one smoothing). However, usually non-linear smoothing techniques are used. See [19] for an overview of various smoothing techniques. In addition to that, it is possible to interpolate the probability of unknown n-gram by combining the probabilities of two or more fitting n-grams of lower order.

Furthermore, new words may appear in the test data that have not been observed in the training data before. Of course, these words make it possible to build completely new n-grams which would all be assigned a probability of 0. To avoid this problem, an artificial token is usually added to the training corpus which represents all unknown tokens and which is assigned a certain probability. In the testing phase, all previously unseen tokens are mapped to this special token.

2.2. Evaluation of the Perplexity

The perplexity is used to evaluate how well an n-gram model fits the test data. The lower the perplexity, the better the test data can be predicted by the model. For a sequence of words $S = (w_1, \dots, w_k)$ of test data, the perplexity is calculated by

$$PPL(S) = P(S)^{-\frac{1}{k}} \tag{2}$$

The perplexity can also be seen as the weighted average branching factor of the data, that is the average number of possible next words that may follow a randomly chosen word of the test data. This interpretation makes the perplexity easy to understand since it is not difficult to see that few possible successors of a word lead to a language of little diversity and, by definition, a low perplexity.

3. Data

Like [16], we use the recordings from the Pitt Corpus[20], which is a part of the DementiaBank[1], as database for our study. It contains recordings from 292 participants who were asked to describe a picture showing a kitchen scene. From these 292 participants 194 suffered from some sort of dementia and 98 healthy speakers serve as control group. Furthermore, the speakers had to be at least 44 years old and must have an initial MMSE score of 10, at least 7 years of education and no history of disorders of the nervous system.

Some of the speakers contributed several recording sessions (at most 7). Moreover, for some sessions the patient's MMSE score at the time of recording is available, too. Unfortunately, there are recordings without a corresponding MMSE score and vice versa.

Like [16], with regard to the dementia group, we narrow our selection to those 168 demented speakers who were diagnosed with AD or probable AD and exclude the speakers who suffered from a different type of dementia (26 speakers). From the latter

[1] http://talkbank.org/DementiaBank/

group, we obtained a total of 255 recordings. Together with the 244 recordings from the control group, a total of 499 recordings were used.

For the actual analysis, we use the transcriptions of the audio file that are available for every recording. Fillers (e.g. *uhm, uhh*) were kept to obtain a more accurate copy of the actual recording. In particular, the transcriptions contain all repetitions, paraphrases, grammatical mistakes, and requests uttered by the participant. Annotations which are not directly linked to the utterances of the test subject (e.g. *clears throat*) are removed. For sentence splitting, Stanford CoreNLP[21] is used.

For the evaluation of the correlation between the MMSE score and the perplexity, we used two different approaches. First, the correlation between the pure values was assessed. For that purpose, all recordings for which a corresponding MMSE score is available are used. These are 234 recording from 166 AD and 181 recordings from 94 healthy speakers.

To take into account longitudinal changes, we additionally measure the correlation on only those recordings, for which at least one additional recording of the same speaker is available[2]. Thus, in the latter set we include 57 speakers with AD and 58 speakers from the control group who contribute a total of 125 respectively 145 pairs of recording and MMSE score.

4. Methods

The SRILM-toolkit[22] is used for the generation of n-grams from the transcriptions that fulfill the above stated criteria. Furthermore, it is used for the calculation of the perplexity. Trigram models are generated and Witten-Bell smoothing is applied. The latter is used since it requires comparatively little input.

As a first step, two trigram models are created from the data: one from all Alzheimer patients, $\mathcal{M}_{\text{alzheimer}}$ and one from all control subjects, $\mathcal{M}_{\text{control}}$. Thus, these models represent the *typical* Alzheimer and healthy speech as it can be inferred from the available data. However, for evaluating the personal speech of each participant, additional models have to be created since the same recording obviously must not appear in both the training as well as the test data. With regard to the speakers that participated in more than one recording session, it is preferable to exclude all their recordings from the training data even when testing on only one of them. That is reasonable since their contributed recordings might contain similar verbalizations or even reoccurring phrases that are typical for the respective speaker but not the whole group and consequently might distort the perplexity.

Hence, we chose a cross-validation approach. For each subject s from the Alzheimer group, a trigram model $\mathcal{M}_{-s}$ is created that includes all recordings from the Alzheimer group other than that obtained from s. Subsequently, the perplexity p_{own} of every speech of the subject is calculated using $\mathcal{M}_{-s}$. Moreover, the perplexity p_{other} is obtained using the model $\mathcal{M}_{\text{control}}$ which is assured not to contain any recordings of s by definition. This process is repeated for all speakers t of the control group. Obviously, $\mathcal{M}_{-t}$ is created using all recordings from the control group except those of t. p_{other} is obtained on the model $\mathcal{M}_{\text{alzheimer}}$.

In addition to that, the difference of perplexities p_{diff} is added as a third feature. To yield comparable results for both groups, it is defined as

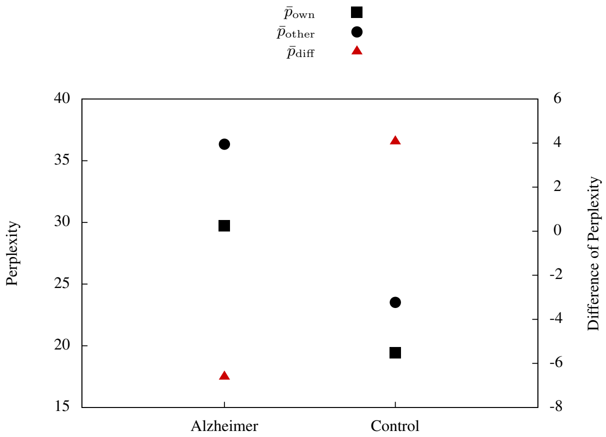

Figure 1: *Mean Values of Perplexity Obtained from Different Models*

$$p_{\text{diff}} = \begin{cases} p_{\text{own}} - p_{\text{other}} & \text{if } s \in \text{AD Group} \\ p_{\text{other}} - p_{\text{own}} & \text{if } s \in \text{Control Group} \end{cases} \tag{3}$$

Consequently, for each recording of both groups a 3-tuple $r = (p_{\text{own}}, p_{\text{other}}, p_{\text{diff}})$ is obtained.

The correlations are assessed in different ways. First, the correlation between the pure MMSE scores and perplexities are calculated on the recordings that fulfill the criteria stated in section 3. Furthermore, the gradients between consecutive recordings of the same speaker, the gradients between all paired recordings of the same speaker and the overall trend of a speaker, obtained by performing a linear regression, are correlated. We consider both Pearson's correlation coefficient r as well as Spearman's rank correlation coefficient ρ.

5. Results

For both the control and the dementia group, the mean values $\bar{p}_{\text{own}}$, $\bar{p}_{\text{other}}$, $\bar{p}_{\text{diff}}$ are calculated. They are visualized in figure 1. It is clearly visible that the mean perplexities $\bar{p}_{\text{own}} = 29.73$ and $\bar{p}_{\text{other}} = 36.34$ of the Alzheimer group are greater than the respective values of the control group which average to $\bar{p}_{\text{own}} = 19.45$ and $\bar{p}_{\text{other}} = 23.52$.

Moreover, it is easy to see that in both cases the perplexity is lower when it is evaluated using the model of the own group than that of the respective other group, i.e. $\bar{p}_{\text{own}} < \bar{p}_{\text{other}}$. This distance is larger for the Alzheimer group than it is for the con-

Table 1: *Correlations Between MMSE Scores and* p_{diff}

		r	ρ
AD	Pure Values	0.367	0.496
	Gradients (Consecutive)	0.26	0.274
	Gradients (Pairwise)	0.235	0.242
	Regression Slope	0.237	0.239
Control	Pure Values	0.180	0.127
	Gradients (Consecutive)	0.095	0.013
	Gradients (Pairwise)	0.151	0.088
	Regression Slope	-0.052	0.046

[2] Note that the sampling is further limited by the little availability of corresponding MMSE scores

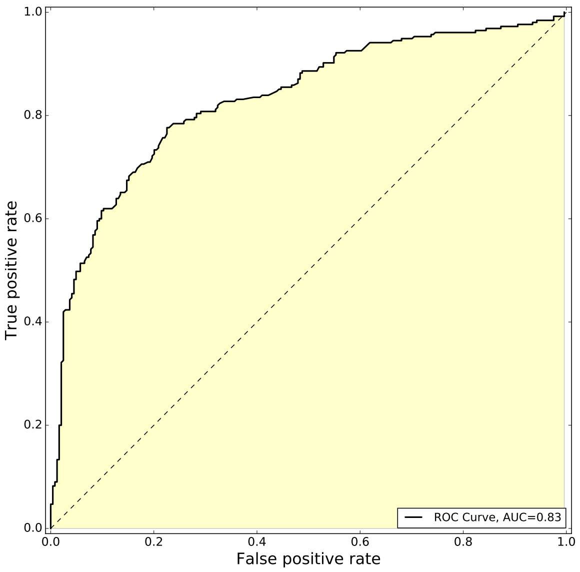

Figure 2: *ROC Curve of binary classifier using $\bar{p}_{\text{diff}}$ as threshold*

trol group. Hence, the variable $\bar{p}_{\text{diff}}$ shows a relatively wide spread.

Using $\bar{p}_{\text{own}}$, the equal error rate is calculated. With $\bar{p}_{\text{own}} = 20.2$, 91 recordings from each group are misclassified. This equals an accuracy of 63.5%. When using $\bar{p}_{\text{other}} = 23$, the accuracy decreases because 102 recordings of each group are misclassified in this case. Thus, the accuracy decreases to 59.1%. However, when using $\bar{p}_{\text{diff}}$, for the evaluation of the error rate, we find threshold of $\bar{p}_{\text{diff}} = -1.41$ since in this case, only 57 Alzheimer recordings and 57 control subjects' recordings are misclassified. This equals an accuracy of 77.1%.

Figure 2 shows the ROC curve of the binary classifier where $\bar{p}_{\text{diff}} = 1.41$ was used as threshold. It indicates, for example, that an Alzheimer patient can correctly be classified with an accuracy of 0.6 while only around 10% of control subjects were wrongly diagnosed with Alzheimer's. The AUC is 0.83.

The MMSE values behave as expected. For the control group, we obtain a mean value $\mu_{\text{control}} = 29.1$ which is close to the maximum score of 30 and is relatively stable ($\sigma_{\text{control}} = 1.17$). In contrary to that, the mean MMSE score of the AD group is expectedly lower ($\mu_{\text{AD}} = 18.54$) and more spread ($\sigma_{\text{AD}} = 5.11$).

The results of the measured correlations between the MMSE score and p_{diff} are presented in table 1. It is easy to see that the correlation is higher for the AD group, regardless of the applied method. Moreover, the rank correlation ρ is markedly lower than r in case of the control group while it is of similar value in case of the AD group.

With exception of the regression slope, the correlations are all significant ($p < 0.5$) in case of the AD group. In contrary to that, none of the correlations are significant in case of the control group.

6. Discussion

As described in section 2.2, a low perplexity indicates that a given text can be well predicted by an n-gram model trained from a different set of texts. In the here described setting, this is the case if the texts of both the training and test data set are describing the happening in a similar way and are coherent. Since the picture shows a precise scene, the vocabulary needed for the description comes from a limited domain, which should yield rather similar n-grams.

The Alzheimer patients, however, tend to describe the scene in an unforeseen way and often divert from the actual task. Moreover, they frequently stumble and repeat what they had uttered previously using different formulations. For example, the speaker of recording 018-0[3] starts to talk about his difficulties with the task at the end of the description. The speaker of 046-0 states that the mother broke one of the plates which is obviously not observable in the picture. The speaker of recording 144-1 sprinkles the description with some narrations from her home and family. The speaker of 551-0 first describes that "she" is washing dishes and later restate the fact that there is a female person, "a lady" in the room who is doing the dishes.

In addition to that, their descriptions often contain incomplete phrases or interruptions. Frequently, patients do not remember the correct word for the object they perceive. Consequently, those objects are often referred to as "thing" or "something". Considering 343-0 the description of the boy falling off the stool is interrupted by a query about the word of the object "stool" and then it is denoted as "thing". Incomplete phrases can be observed in 368-0 who only raises some subjects without giving information. She utters: "and the window, and the grass and stuff outside. I don't know where everything...".

These quoted examples correspond to findings from literature as described in section 1.1. Of course, not all of the mentioned characteristics are present in the speech of every Alzheimer patient and some of the characteristics can be found in many more speakers than listed here. Although the given examples illustrate the peculiarities of the patients' language, they do not fully cover all facets of it since the actual usage strongly varies between the speakers. However, they illustrate how the presence and combination of those features lead to a very unique and unpredictable speech.

Considering the measured correlations, the high discrepancy between the rank correlation coefficient ρ and the Pearson correlation p in case of the control group indicate that the measurements obtained from this group are not normally distributed. Moreover, since all obtained correlations from this group are of no significance, it likely that the correlations were obtained by chance. This is of little surprise since the MMSE score should be stable for healthy adults and arrange close the maximum value of 30.

This is contrasted with the overall higher agreement between r and ρ in case of the AD group. Here, the MMSE score is supposed to decrease after the onset of AD. The observable correlations between the MMSE score and the perplexity partly reveals this trend which is further assured by the significance of the correlations.

7. Acknowledgements

The work reported here was carried out during the 2016 Jelinek Memorial Summer Workshop on Speech and Language Technologies, which was supported by Johns Hopkins University via DARPA LORELEI Contract No HR0011-15-2-0027, and gifts from Microsoft, Amazon, Google, and Facebook.

[3] see http://talkbank.org/browser/index.php?url=DementiaBank /English/Pitt/Dementia/cookie/<recordingID>.cha for this and the following examples

8. References

[1] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders*, 4th ed. Washington, DC: American Psychiatric Press, 1994.

[2] M. Prince, R. Bryce, E. Albanese, A. Wimo, W. Ribeiro, and C. P. Ferri, "The global prevalence of dementia: A systematic review and metaanalysis," *Alzheimer's & Dementia*, vol. 9, no. 1, pp. 63 – 75.e2, 2013.

[3] A. Husband and W. Alan, "Different types of dementia," *The Pharmaceutical Journal*, vol. 277, pp. 579–582, November 2006.

[4] D. S. Geldmacher and P. J. Whitehouse, "Evaluation of dementia," *New England Journal of Medicine*, vol. 335, no. 5, pp. 330–336, 1996.

[5] M. Prince, A. Wimo, M. Guerchet, G. Ali, Y. Wu, and M. Prina, "World alzheimer report 2015. The global impact of dementia. An analysis of prevalence, incidence, cost & trends," Alzheimer's Disease International, London, Tech. Rep., 2015.

[6] J. T. O'Brien, "Role of imaging techniques in the diagnosis of dementia," *The British Journal of Radiology*, vol. 80, no. special_issue_2, pp. 71–77, 2007.

[7] 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.

[8] R. T. Woods, "Discovering the person with Alzheimer's disease: cognitive, emotional and behavioural aspects," *Aging & Mental Health*, vol. 5, pp. 7 – 16, 2001.

[9] D. Kempler, "Language changes in dementia of the Alzheimer type," in *Dementia and Communication*, R. Lubinski, Ed. Philadelphia: B.C. Decker, 1991.

[10] A.-L. R. Adlam, S. Bozeat, R. Arnold, P. Watson, and J. R. Hodges, "Semantic knowledge in mild cognitive impairment and mild alzheimer's disease," *Cortex*, vol. 42, no. 5, pp. 675–684, 2006.

[11] H. S. Kirshner, "Primary progressive aphasia and Alzheimer's disease: Brief history, recent evidence," *Current Neurology and Neuroscience Reports*, vol. 12, no. 6, pp. 709–714, 2012.

[12] 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.

[13] D. N. Ripich and B. Y. Terrell, "Patterns of discourse cohesion and coherence in Alzheimer's disease," *Journal of Speech and Hearing Disorders*, vol. 53, no. 1, pp. 8–15, 1988.

[14] K. Dijkstra, M. S. Bourgeois, R. S. Allen, and L. D. Burgio, "Conversational coherence: Discourse analysis of older adults with and without dementia," *Journal of Neurolinguistics*, vol. 17, no. 4, pp. 263–283, 2004.

[15] J. Weiner and T. Schultz, "Detection of intra-personal development of cognitive impairment from conversational speech," in *Speech Communication; 12. ITG Symposium*, 2016, pp. 240–244.

[16] 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.

[17] D. Jurafsky and J. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, ser. Practical Resources for the Mental Health Professionals Series. Prentice Hall, 2000.

[18] C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[19] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–393, 1999.

[20] 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.

[21] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Association for Computational Linguistics (ACL) System Demonstrations*, Baltimore, 2014, pp. 55–60.

[22] A. Stolcke, "Srilm-an extensible language modeling toolkit." in *Proc. Intl. Conf. on Spoken Language Processing*, vol. 2, Denver, 2002, pp. 901–904.