Speech-Based Detection of Alzheimer's Disease in Conversational German

Jochen Weiner, Christian Herff, Tanja Schultz

Cognitive Systems Lab, University of Bremen, Germany

jochen.weiner@uni-bremen.de

Abstract

The worldwide population is aging. With a larger population of elderly people, the numbers of people affected by cognitive impairment such as Alzheimer's disease are growing. Unfortunately, there is no known cure for Alzheimer's disease. The only way to alleviate it's serious effects is to start therapy very early before the disease has wrought too much irreversible damage. Current diagnostic procedures are neither cost nor time efficient and therefore do not meet the demands for frequent mass screening required to mitigate the consequences of cognitive impairments on the global scale.

We present an experiment to detect Alzheimer's disease using spontaneous conversational speech. The speech data was recorded during biographic interviews in the Interdisciplinary Longitudinal Study on Adult Development and Aging (ILSE), a large data resource on healthy and satisfying aging in middle adulthood and later life in Germany. From these recordings we extract ten speech-based features using voice activity detection and transcriptions. In an experimental setup with 98 data samples we train a linear discriminant analysis classifier to distinguish subjects with Alzheimer's disease from the control group. This setup results in an F-score of 0.8 for the detection of Alzheimer's disease, clearly showing our approach detects dementia well.

Index Terms: computational paralinguistics, ILSE, dementia, Alzheimer

1. Introduction

The worldwide population is aging rapidly. While in 1950 about 8 % of the worldwide population was 60 years or older, that number had risen to 12 % in 2013 and is expected to rise to 21 % by 2050 [1, p. 11]. With this increase in the elderly population cognitive decline in the elderly becomes a major challenge for the aging societies. Depending on the world region, dementia has a prevalence between 4 % and 8 % in the 60+ age group, with 46.8 million people affected and health care costs of over US$ 818 billion in 2015 [2]. On a more individual level, the disease has major impacts on the lives of those affected by it, their relatives and their caregivers. With up to 70 % of the cases, Alzheimer's disease is the most common form of dementia [3]. While there is no known cure for this form of dementia, its effects can be delayed if therapy starts early [2]. However, therapy can only start early if the disease is diagnosed early which requires frequent examinations of the cognitive status of the elderly age group.

Examinations typically include a large number of neuropsychological tests such as the Mini Mental State Examination [4]. The results of these test are then used by psychiatrists to make a diagnosis. While this procedure produces reliable results, it is also very time consuming and expensive. In order to enable longitudinal cognitive status monitoring on a large scale, a method for diagnosing the disease needs to be found that is fast and cheap and can be made widely available.

It has been shown for both Alzheimer's disease [5] and dementia of non-Alzheimer type [6] that dementia affects human speech and language. Furthermore, the significant changes to speech and language use caused by dementia occur very early in the course of the disease [7]. Speech is therefore a promising candidate as a source of information for new approaches to diagnosing dementia.

In recent years, features extracted from speech have already been used to automatically detect mild cognitive impairment (MCI) and Alzheimer's disease (AD). The subjects in most of these studies are patients who are already being treated for their cognitive impairment and suitable healthy controls. The subjects' speech is recorded while they carry out a subset of the same neuropsychological tests that psychiatrists use for their diagnosis of cognitive impairment. Features for the detection of the cognitive impairment are then extracted from these recordings. Under these carefully controlled conditions good classification results have been reported using acoustic or prosodic features [8, 9, 10, 11], and linguistic or text-based features [12, 13, 14].

Since these studies require their subjects to carry out neuropsychological tests, they provide little improvement over the current diagnosis procedure in which psychiatrists analyze the results of the tests. Using spontaneous speech, e.g. from interviews, requires less preparation, can be recorded by a person with minimal diagnostic knowledge and may provide subjects with a more comfortable situation. This form of less controlled speech has received less attention, but classification results on par with the test recordings have been reported [15, 16].

All these studies followed a cross-sectional approach in which each subject was recorded and examined exactly once. Longitudinal corpora in which each subject is recorded and examined multiple times contains more information such as individual aging and progression of disease in individuals. Furthermore, these corpora enable research closer to the motivational idea that frequent longitudinal examinations of the elderly are required to mitigate the consequences of cognitive impairment. Recordings of neuropsychological tests from a four-year longitudinal study were used by Yu et al. [17, 18]. They did not work on intra-personal longitudinal detection of changes from cognitively healthy to cognitive impaired, but detected the current cognitive status like the previous studies. The difference to the studies mentioned previously is that one person could contribute to both classes, cognitively healthy and cognitively impaired, if their cognitive diagnosis changed over the course of the study.

In this paper we present our baseline work on speech-based features for the detection of aging-associated cognitive decline (AACD) and Alzheimer's disease (AD) using spontaneous conversational speech. A very pronounced difference between people with dementia and healthy controls that has been reported is

that people suffering from dementia tend to hesitate more often and make longer pauses [16]. We therefore use features based on the occurence and duration of pauses to differentiate between healthy controls and subjects affected by cognitive decline.

For our experiments we use data from the *Interdisciplinary Longitudinal Study on Adult Development and Aging* (ILSE) [19]. ILSE is a collection of data on healthy and satisfying aging in middle adulthood and later life, not a database tailored specifically for the automatic detection of dementia. A group of 1,000 participants was recruited through community registers so that this group is representative for the sampled population: members of two birth cohorts that lived in two urban centers in Germany. The elder cohort were around 60 years old when the study began, while the younger cohort were around 40 years old. Over the course of more than 20 years these participants have contributed data to ILSE in four measurements. While the participants were very young to be examined for cognitive impairment when the study began, they have now reached the age where dementia occurs. The prevalence of dementia in each age group at each measurement [19] is in the range expected for that age group in Germany [2, p. 20]. This means that the data in ILSE provides us with a scenario that is very close to the conditions we would find in real longitudinal cognitive status monitoring in this country.

This paper is organized as follows: In Section 2 we go into detail about the ILSE data. After the data we describe the features (Section 3) that we used to perform the experiment (Section 4) and obtain the results (Section 5). We conclude and take a look at our future work in Section 6.

2. Database

ILSE comprises data to assess the participants' personality, cognitive functioning, subjective well-being, and health: results from psychological, cognitive, physical, and dental examinations, as well as biographic information derived from questionnaires and semi-standardized biographic interviews. Specific to the detection of dementia from speech, ILSE contains recordings of over 8,000 hours of biographic interviews as well as cognitive diagnoses for the participants. Interviewers led through the interviews with short prepared questions and encouraged the participants to answer in detail. Thus the largest portion of the recordings is spontaneous participant speech. Cognitive diagnoses were made by psychiatrists using a range of neuropsychological, anamnestic, clinical, and laboratory tests.

The participants were invited to participate in four data collection periods, and one measurement was taken from every participant in each period. However, only 90 % of the participants returned for the second measurement and only 65 % participated in the third measurement, which means less data is available for the later measurements.

At the beginning of the study, all the participants were cognitively healthy or so young that cognitive impairment was so unlikely that they were not tested. Over time, some participants developed cognitve impairment. This means ILSE is a real world data set in which the cognitive diagnoses are distributed according to the natural prevalence. Due to this natural influence the cognitive diagnoses in ILSE form a highly unbalanced data set.

For this work we used a subset of the interviews that fulfill the following constraints:

- Availability of cognitive diagnosis: The fourth measurement is currently being recorded and cognitive diagnoses have not yet been finalized for this measurement. Therefore we only take into account data from the first three measurements for which cognitive diagnoses are available.
- Availability of transcriptions: A small portion of 384 hours of interview recordings from the first three measurements has been transcribed manually. Since some of our features (see Section 3) require a transcription, we restrict the dataset to the interviews that have been transcribed.

The data that we selected according to these constraints was speech data recorded from 74 participants. We do not consider the longitudinal character of the data, but treat each measurement independently from the others. Thus one participant may e.g. contribute two measurements with a control diagnosis and one measurement with an AACD diagnosis which we treat as three separate samples. Table 1 shows the samples that we use with their cognitive diagnoses. In these samples, the diagnoses are highly unbalanced. However, the prevalence of Alzheimer's disease is in the range to be expected in the age group of the 70 to 74-year-olds in Germany [2, p. 20]. Therefore this dataset provides a realistic scenario for the detection of Alzheimer's disease.

		Diagnoses		
		control	**AACD**	**AD**
Measurement	**1**	51	4	-
	2	19	8	-
	3	10	1	5
Total		80	13	5

Table 1: Cognitive diagnoses from the three ILSE measurements that meet the limitation factors.

The interviews have not been manually segmented at speaker turns and there exists no manual alignment between audio recording and transcription. We therefore employed a long audio alignment procedure [19] to align the transcriptions with the audio recordings and used these alignments to create a speaker segmentation. From this segmentation we selected all the segments with participants' speech, a total of 230 hours of audio recordings.

3. Features

Manual linguistic analyses have already shown that detection of cognitive impairment using recordings of the ILSE interviews is possible [20]. The features for the detection of dementia we employ in this work do not use linguistic information but focus on features based on acoustic information from the participants' speech. They were automatically extracted from the interview recordings and their manual transcriptions. Our features are inspired by the prosodic features presented by Khodabakhsh et al. [16] which, in turn, are similar to the features used in most of the related work.

The majority of the features is created from segments of silence and segments of speech which we obtained from voice activity detection (VAD). Our VAD system uses a Hidden-Markov-Model recognizer in our in-house toolkit BioKIT [21]. The recognizer has two models, one for silence and one for non-silence, both of which are modeled by Gaussian-Mixture-

Models using 128 Gaussians each. Given Mel-frequency cepstral coefficients (MFCCs) with first and second order derivatives plus zero crossing rate as input, the VAD performs a two-pass recognition process with Maximum Likelihood Linear Regression (MLLR) adaptation between the two passes. The resulting partition is then smoothed so that the duration of a silence segment is at least 0.2 seconds.

Given the partition of speaker turns into silence segments and speech segment from the VAD, we extract six acoustic features based on silence segments described in Section 3.1. The partition from the VAD together with the manual transcriptions enable us to extract four features that use acoustic information and textual information which we describe in Section 3.2. After extraction these ten features are stacked together into one feature vector for each interview.

3.1. Acoustic features based on silence

Acoustic features have also been called vocal features in the literature. Purely acoustic features are those that do not capture what was said or how information was transferred using speech, but instead capture how the words were uttered. Our acoustic features based on silence use the silence and speech segments found by the VAD to capture the occurrence and duration of pauses.

mean silence duration
: The mean duration of a silence segment should show differences between participants with cognitive impairment and control subjects, if dementia causes longer pauses.

mean speech duration
: The mean speech duration is the mean duration of a speech segment. If people suffering from dementia make more pauses, speech will more often be interrupted and each speech segment will have to be shorter.

silence rate
: Silence rate is calculated by dividing the total duration of the silence segments by the duration of the speaker turns. It combines the idea behind the first two features: If people suffering from dementia make more and longer pauses, then the portion of silence in their speech will be higher.

silence count ratio
: For the silence count ratio the number of silence segments is divided by the total number of segments. In the VAD a silence segment is always followed by a speech segment. The idea behind this feature is to measure the silence at the beginning of a speaker turn (hesitant reply) and at the end of a speaker turn (trailing open end to last speech segment).

silence-to-speech ratio
: The silence-to-speech ratio is the number of silence segments divided by the number of speech segments. This ratio is a measure of the hesitation rate.

mean silence count
: The mean silence count is calculated as the number of silence segments divided by the duration of the speaker turns. The first two features try to capture whether people suffering from dementia make more pauses. This feature is a third expression trying to capture the same information.

3.2. Acoustic features based on silence and transcriptions

In addition to the purely acoustic features we also extract features that combine information from the VAD partition with information from the transcription of the speech. These acoustic features based on silence and transcriptions take into account which and how many words were uttered. Thus they can capture some information about the spoken text. For this work we relied on manual transcriptions but these can later be replaced by transcriptions produced by automatic speech recognition.

silence-to-word ratio
: The silence-to-word ratio is the number of silence segments divided by the transcription word count. It is a measure of the hesitation rate that accommodates for different speaking rates.

long-silence-to-word ratio
: Long silences are those silences that are longer than one second. The long-silence-to-word ratio is the number of such long silence segments divided by the transcription word count. Long silence may have a different meaning in speech than shorter ones. This feature measures the rate of long pauses while taking into account different speaking rates.

word rate
: The word rate describes the speaking rate at the word level. It is calculated as the total number of spoken words divided by the duration of the speaker turns.

phoneme rate
: The phoneme rate measures the speaking rate at the phoneme level. This rate is the total number of spoken phonemes divided by the duration of the speaker turns. The phoneme sequence required to extract this feature is generated by translating the word transcription to a phoneme sequence using a pronunciation dictionary.

4. Experiment

We conducted a classification experiment with three classes: *control* for cognitively healthy participants, *AACD* for participants with early-stage cognitive decline, and *AD* for participants with progressed cognitive decline. We extracted the ten features described in Section 3 for the dataset of 98 interviews described in Section 2. The ten features were then combined into one feature vector for each interview. This feature extraction results in 98 10-dimensional feature vectors with the labels shown in Table 2.

	control	**AACD**	**AD**
Samples	80	13	5

Table 2: The number of samples in each class. Also see Table 1.

We use these features to train a classifier to detect AD and AACD. Since the classes are unbalanced, we did not split the data into training and test sets. Instead we trained and evaluated our models in a cross-validation. As there are only 5 samples in the *AD* class we chose a stratified 3-fold cross-validation. Using more folds or non-stratified cross-validation would mean that the data distribution in the folds would no longer match that of the data and lead to meaningless classification results.

In the cross-validation we trained linear discriminant analysis (LDA) classifiers with singular value decomposition and no

shrinkage. The classifiers are trained for the multiclass classification problem of the three classes *control*, *AACD* and *AD*. For the cross-validation and LDA we used the implementations in scikit-learn [22].

5. Results

The results of the experiment (Section 4) are depicted in the confusion matrix in Figure 1. Precision, recall and F-score for the three classes are given in Table 3.

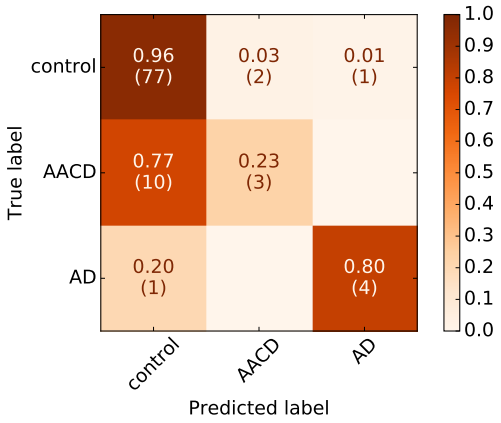

Figure 1: Normalized confusion matrix of the result of the three-class-classification of the samples described in Table 2 (The actual number of samples is given in parenthesis).

	precision	**recall**	**F-score**
control	0.88	0.96	0.92
AACD	0.60	0.23	0.33
AD	0.80	0.80	0.80

Table 3: Precision, recall and F-score of the three-class-classification of the samples described in Table 2.

The classes *control* and *AD* are distinguished well. The F-score for the class *AD* is 0.8, with recall of 0.8 and precision of 0.8 (Table 3). In absolute numbers, four of the five samples for class *AD* were recognized correctly while one was incorrectly labeled *control* and one *control* label was incorrectly labeled *AD*. To interpret this result we compared this F-score to the highest naively constructible F-score, which is obtained by assigning all samples to the *AD* class and has a value of 0.1. This comparison shows that the LDA classifier can detect *AD* much better than any naive classification.

The class *AACD* could clearly be distinguished from the class *AD* with no confusion between the two classes. However, it was very often confused with the *control* class, with over two thirds of the *AACD* samples labeled as *control* and low values for precision, recall and F-score (Table 3). Unfortunately this means that our classifier is not yet able to detect early-stage cognitive decline. Distinguishing between these two classes will be a major objective for our future work so that we will move towards very early detection of cognitive decline.

The overall accuracy of our classifier is 85.7 %, the unweighted average recall is 0.66. These results show clearly that people suffering from Alzheimer's disease can be distinguished from healthy controls using the spontaneous conversational speech in the ILSE corpus. On the other hand, our classifier did not find any distinguishable differences between the healthy controls and people suffering from AACD.

6. Conclusions and Future Work

As the population is aging and the numbers of people affected by cognitive impairment are increasing, cost and time effective methods to diagnose cognitive impairment such as Alzheimer's disease need to be found. Speech is among the first cognitive domains affected by AD and is therefore a promising candidate in the search for new diagnostic methods.

We presented automatic detection of AD and AACD in recordings of spontaneous conversational speech from interviews in the ILSE corpus. This longitudinal corpus contains a representative sample of two age-groups in Germany. The prevalence of dementia in this corpus is in the range of the expected prevalence in the whole population. Therefore the data are well suited to investigate new diagnostic methods.

For the detection of AD and AACD we used ten features automatically extracted from the interview recordings. Six of these features are purely acoustic features that rely solely on voice activity detection (VAD) while the remaining four features are based on VAD and textual transcriptions. In our experiment, an LDA classifier achieved 85.7 % accuracy for the recognition of the three classes control, AACD and AD. The F-score for Alzheimer's disease is 0.8, which is a clear demonstration that speech-based features from the spontaneous conversational ILSE interviews are well suited for the investigation of novel methods for the diagnosis of dementia using speech. These results demonstrate for the first time that dementia can be recognized in spontaneous conversational German with a quality par with results published for other languages and other types of recordings.

This is the result of a cross-sectional experiment across three of the measurements in ILSE. In this work we have not investigated the longitudinal character of the data provided by ILSE. This aspect will be part of our future work in which we will investigate the detection of changes in the cognitive diagnoses in individual participants over time.

The features used in this work rely on the availability of transcriptions. In the future, we plan to perform unsupervised speaker segmentation to find participant speaker turns and apply automatic speech recognition to provide transcriptions. Together with our classifier dementia can then be detected fully automatically.

7. Acknowledgements

We thank Claudia Frankenberg, Britta Wendelstein and Johannes Schröder at the University of Heidelberg and the University Hospital Heidelberg, Germany for providing us with the data that enabled the research presented in this paper.

8. References

[1] United Nations, Department of Economic and Social Affairs, Population Division, *World Population Ageing 2013*. United Nations, 2013.

[2] M. Prince, A. Wimo, M. Guerchet, G.-C. Ali, Y.-T. Wu, and M. Prina, *World Alzheimer Report 2015. The Global Impact of Dementia: an Analysis of Prevalence, Incidence, Cost and Trends*. London: Alzheimer's Disease International, 2015.

[3] World Health Organization and Alzheimers Disease International,

Dementia: a public health priority. World Health Organization, 2012.

[4] 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.

[5] J. Appell, A. Kertesz, and M. Fisman, "A study of language functioning in Alzheimer patients," *Brain and language*, vol. 17, no. 1, pp. 73–91, 1982.

[6] J. Reilly, A. D. Rodriguez, M. Lamy, and J. Neils-Strunjas, "Cognition, language, and clinical pathological features of non-alzheimer's dementias: An overview," *Journal of Communication Disorders*, vol. 43, no. 5, pp. 438 – 452, 2010.

[7] R. 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.

[8] A. Satt, A. Sorin, O. Toledo-Ronen, O. Barkan, I. Kompatsiaris, A. Kokonozi, and M. Tsolaki, "Evaluation of speech-based protocol for detection of early-stage dementia." in *INTERSPEECH 2013 – 14th Annual Conference of the International Speech Communication Association*, 2013, pp. 1692–1696.

[9] A. Satt, R. Hoory, A. König, P. Aalten, and P. H. Robert, "Speech-based automatic and robust detection of very early dementia." in *INTERSPEECH 2014 – 15th Annual Conference of the International Speech Communication Association*, 2014, pp. 2538–2542.

[10] F. Espinoza-Cuadros, M. A. Garcia-Zamora, D. Torres-Boza, C. A. Ferrer-Riesgo, A. Montero-Benavides, E. Gonzalez-Moreira, and L. A. Hernandez-Gómez, "A spoken language database for research on moderate cognitive impairment: design and preliminary analysis," in *Advances in Speech and Language Technologies for Iberian Languages*. Springer, 2014, pp. 219–228.

[11] L. Tóth, G. Gosztolya, V. Vincze, I. Hoffmann, and G. Szatlóczki, "Automatic detection of mild cognitive impairment from spontaneous speech using asr," in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association*, 2015, pp. 2694–2698.

[12] D. Hakkani-Tür, D. Vergyri, and G. Tür, "Speech-based automated cognitive status assessment." in *INTERSPEECH 2010 – 11th Annual Conference of the International Speech Communication Association*, 2010, pp. 258–261.

[13] E. T. Prud'hommeaux and B. Roark, "Extraction of narrative recall patterns for neuropsychological assessment." in *INTERSPEECH 2011 – 12th Annual Conference of the International Speech Communication Association*, 2011, pp. 3021–3024.

[14] M. Lehr, E. T. Prud'hommeaux, I. Shafran, and B. Roark, "Fully automated neuropsychological assessment for detecting mild cognitive impairment." in *INTERSPEECH 2012 – 13th Annual Conference of the International Speech Communication Association*, 2012, pp. 1039–1042.

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

[16] A. Khodabakhsh, F. Yesil, E. Guner, and C. Demiroglu, "Evaluation of linguistic and prosodic features for detection of Alzheimer's disease in Turkish conversational speech," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–15, 2015.

[17] B. Yu, T. F. Quatieri, J. R. Williamson, and J. C. Mundt, "Prediction of cognitive performance in an animal fluency task based on rate and articulatory markers." in *INTERSPEECH 2014 – 15th Annual Conference of the International Speech Communication Association*, 2014, pp. 1038–1042.

[18] ——, "Cognitive impairment prediction in the elderly based on vocal biomarkers," in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association*, 2015, pp. 3734–3738.

[19] J. Weiner, C. Frankenberg, D. Telaar, B. Wendelstein, J. Schröder, and T. Schultz, "Towards Automatic Transcription of ILSE – an Interdisciplinary Longitudinal Study of Adult Development and Aging," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016.

[20] B. Wendelstein, *Gesprochene Sprache im Vorfeld der Alzheimer-Demenz. Linguistische Analysen im Verlauf von prklinischen Stadien bis zur leichten Demenz*. Heidelberg: Winter, 2016.

[21] D. Telaar, M. Wand, D. Gehrig, F. Putze, C. Amma, D. Heger, N. T. Vu, M. Erhardt, T. Schlippe, M. Janke, C. Herff, and T. Schultz, "BioKIT - Real-time decoder for biosignal processing," in *INTERSPEECH 2014 – 15th Annual Conference of the International Speech Communication Association*, 2014, pp. 2650–2654.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.