Detection of Intra-Personal Development of Cognitive Impairment From Conversational Speech

Jochen Weiner, Tanja Schultz

Cognitive Systems Lab, Universität Bremen, 28359 Bremen, Germany
Email: jochen.weiner@uni-bremen.de
Web: csl.uni-bremen.de

Abstract

As the population in developed countries is aging, cognitive impairment such as Alzheimer's disease becomes an urging challenge for these societies. In order to mitigate the consequences, diagnosing cognitive impairment early is crucial. We present automatic detection of an intra-personal development of cognitive impairment from speech. Using conversational speech data from the ILSE corpus we detect subjects which were considered cognitively healthy at one examination and were diagnosed with a cognitive impairment at a later examination.

From the speech recordings we extract 14 speech-based features using voice activity detection and transcriptions. With these features we train a linear discriminant analysis classifier that distinguishes subjects who developed a cognitive impairment from subjects who did not. The classifier achieves an accuracy of 80.4%, classifying half the cognitively impaired subjects correctly and assigning that label to hardly any cognitively health subjects. This shows our approach is well suited for longitudinal cognitive status monitoring.

1 Introduction

The population in developed countries is aging rapidly. In Germany, for example, the most populous age group in the year 1950 were ten-year-olds and in 2000 forty-year-olds; for 2050 it is estimated to be the sixty-year-olds [1]. With this increase in the elderly population cognitive decline in the elderly becomes a major challenge for Germany as for any other aging society. In Germany, dementia has a prevalence of 6.8 % in the 60+ age group [2, p. 20]. World-wide, 46.8 million people are affected and health care costs for the treatment of dementia have risen to 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.

Cognitive status 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, vocal or prosodic features [8–11], and linguistic or text-based features [12–14].

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, allows for recording 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 work on the detection of intra-personal development of cognitive impairment: using speech data from two examinations we detect subjects that were cognitively healthy during one examination and cognitively impaired during a later examination. To the best of our knowledge no such work has been published to date. We use spontaneous conversational speech data from the *Interdisciplinary Longitudinal Study on Adult Development and Aging* (ILSE) [19], 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 large-scale longitudinal cognitive status monitoring in this country.

This paper is organized as follows: Section 2 introduces the ILSE data in detail. 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 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 conversational 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. 90 % of the participants returned for the second measurement and 65 % participated in the third measurement, which means less data is available for the later measurements. For this work we used a subset of the interviews that fulfill the following constraints:

- Availability of longitudinal data: Not all participants took part in all measurements. We select the data from participants who participated in at least two consecutive measurements.
- 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 23 participants. The interviews have not been segmented at speaker turns and there exists no 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 112 hours of audio recordings.

In ILSE, there is one diagnosis for cognitively healthy participants (*control*) and four diagnoses for cognitively impaired participants (*aging-associated cognitive decline (AACD)*, *mild cognitive disorder (MCD)*, *Alzheimer's disease (AD)* and *vascular dementia (VAD)*). In this work we investigate the intra-personal development of cognitive impairment over time. This means a change from the control diagnosis at one measurement to one of the four cognitively impaired diagnoses at a later measurement. Table 1 shows how often such a change occurred between two measurements in the selected data.

		Change in diagnosis from healthy to impaired	
		yes	**no**
Measurements	**1 → 2**	8	15
	2 → 3	6	8
	1 → 3	2	12
Total		16	35

Table 1: The number of participants whose cognitive diagnosis changed from cognitively healthy to cognitively impaired between two measurements.

As Table 1 shows, a total of 16 participants developed cognitve impairment between two measurements. These 16 participants were born in 1930-32, so they were 61-66 years old during the first measurement, 65-70 years old during the second measurement and 73-77 years old during the third measurement.

3 Features

We employ automatically extracted features for the detection of dementia that do not use linguistic information but focus on features based on acoustic information from the participants' speech. These 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.

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]. Therefore, 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 [20]. 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 ten acoustic features based on the VAD partition 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 14 features are stacked into one feature vector for each interview.

3.1 Acoustic features from speech

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. These features have also been called vocal or prosodic features in the literature. Our acoustic features from speech use the silence and speech segments found by the VAD to capture the occurrence and duration of pauses.

mean/median/variance silence duration
: The mean, median and variance duration of a silence segment may indicate differences between participants with cognitive impairment and control subjects, if dementia causes longer pauses.

mean/median/variance speech duration
: The mean, median and variance 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 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. It is included here so that we can investigate the most suitable features for the detection.

3.2 Acoustic features from speech 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 the number of words and phonemes that 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 (ASR). Since we use only rough information about the spoken text such as the number of words and phonemes, it is likely that even with erroneuos ASR output the resulting features will closely match those extracted from the transcription.

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, thus compensating for different word lengths. 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 mapping the word transcription to a phoneme sequence using a pronunciation dictionary.

4 Experiment

For our experiment we extracted the features described in Section 3 for each measurement of the dataset described in Section 2. The 14 features were then combined into one feature vector for each interview. To model the intra-personal change between two measurements (Table 1) we subtracted the feature vector of the earlier measurement from the feature vector of the later measurement. We normalized the resulting vector to unit length. The final feature vectors now point into the direction in which the feature vectors changed between the two measurements.

These normalized difference features are the samples we use to train a classifier to detect a change from cognitively healthy to cognitively impaired. We did not split the data into training and test sets but trained and evaluated our models in a stratified 6-fold cross-validation, choosing $k = 6$ for the cross-validation to keep the variance of the classification error in this small dataset low.

In the cross-validation we trained linear discriminant analysis (LDA) classifiers with singular value decomposition and no shrinkage. The classifiers are trained for the binary classification problem *change* $\leftrightarrow$ *no change*. For the

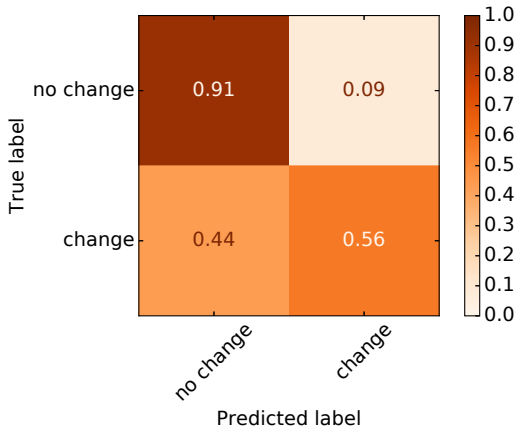

Figure 1: Normalized confusion matrix of the result of the binary classification of the samples described in Table 1.

cross-validation and LDA we used the implementations in scikit-learn [21].

5 Results

The results of the experiment (Section 4) are depicted in the confusion matrix in Figure 1.

These results are promising in the scenario of longitudinal cognitive status monitoring: A recall of 0.91 of the *no change* class means that very few cognitively healthy participants are incorrectly labeled as cognitively impaired. With a precision of 0.56 for the *change* class indicate that over half of the participants with a change in their cognitive diagnosis were correctly identified. The classifier has an F-score of 0.87 for the class *no change* and 0.64 for the class *change*. To interpret this result we compared these F-scores to the highest naively constructible F-scores, which are obtained by assigning all samples to the respective class. The highest naively constructible F-score for the class *no change* has a value of 0.81 and the highest naively constructible F-score for the class *change* is 0.48. This comparison shows that the LDA classifier can detect a *change* in the diagnosis much better than any naive classification. Overall, 41 of the 51 samples were classified correctly, yielding and accuracy of 80.4%. These results show that half of the participants who have undergone a change from cognitively healthy to cognitively impaired can be identified while assigning this diagnosis to hardly any participants who have stayed cognitively healthy.

6 Conclusions

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

We presented the automatic detection of intra-personal development of cognitive impairment 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 and 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 intra-personal development of cognitive impairment we have investigated the longitudinal character of the the data provided by the ILSE corpus. We used 14 features automatically extracted from the interview recordings. Ten 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 80.4 % accuracy for the binary classification problem *change* $\leftrightarrow$ *no change*. Furthermore, the classifier identified over half of the cognitively impaired participants while assigning that diagnosis to hardly any healthy participants, which is a desired result for a future large-scale longitudinal cognitive status monitoring.

7 Acknowledgements

We thank Alan W Black for the valuable discussions that led to this paper. We also 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.

References

[1] Statistische Ämter des Bundes und der Länder, "Bevölkerungs- und Haushaltsentwicklung im Bund und in den Ländern.," *Demografischer Wandel in Deutschland*, vol. 1, 2011.

[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 Alzheimer's 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*, pp. 1692–1696, 2013.

[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*, pp. 2538–2542, 2014.

[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*, pp. 219–228, Springer, 2014.

[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*, pp. 2694–2698, 2015.

[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*, pp. 258–261, 2010.

[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*, pp. 3021–3024, 2011.

[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*, pp. 1039–1042, 2012.

[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, pp. 1569–1574, IEEE, 2005.

[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*, pp. 1038–1042, 2014.

[18] B. Yu, T. F. Quatieri, J. R. Williamson, and J. C. Mundt, "Cognitive impairment prediction in the elderly based on vocal biomarkers," in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association*, pp. 3734–3738, 2015.

[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] 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*, pp. 2650–2654, 2014.

[21] 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.