SPEECH REVEALS FUTURE RISK OF DEVELOPING DEMENTIA: PREDICTIVE DEMENTIA SCREENING FROM BIOGRAPHIC INTERVIEWS

Jochen Weiner¹, Claudia Frankenberg², Johannes Schröder², Tanja Schultz¹

¹Cognitive Systems Lab, University of Bremen, Germany
²Section of Geriatric Psychiatry, University of Heidelberg, Germany

ABSTRACT
Alzheimer’s disease is a progressive incurable condition for which the success of any symptomatic therapy depends crucially on the starting time. Ideally it starts before the disease has caused any cognitive impairments. Our work aims at developing speech-based dementia screening methods that detect dementia as early as possible. Here, we aim to predict the outbreak even before clinical screening tests can diagnose the disease. Using the longitudinal ILSE study, we automatically extract features from biographic interviews and predict the development of dementia 5 and 12 years into the future. Our prediction system achieves results of 73.3% and 75.7% unweighted average recall (UAR), respectively, which clearly outperform a prediction based on prior diagnoses or disease prevalence. Thus, the automated analysis of spoken interviews offers a highly effective prediction procedure that allows for easy-to-use, cost-effective casual testing.

Index Terms— dementia screening, predictive screening, ILSE

1. INTRODUCTION
Alzheimer’s disease is the most common condition that leads to a neuro-degenerative syndrome called dementia [1]. Being diagnosed with Alzheimer’s is bad news for patients: there is no known cure [2] and being a degenerative disease, the patient’s condition will deteriorate over time. However, the course of the disease can be influenced positively through secondary therapy. The success of such a therapy is extremely dependent on the start of the therapy: the sooner therapy can start, the more patients can benefit [3]. Ideally a patient’s need for therapy is detected before the disease has wrought any considerable damage and so therapy can start when the patient’s cognitive abilities have not yet been noticeably impaired.

Dementia affects speech and language from a very early stage and speech and language are strong indicators for dementia [4, 5]. This has been exploited through automatic speech processing [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]. Nearly all research focuses on state screening: screening for the patient’s current cognitive state, i.e. the diagnosis at the time of a screening session using the patient’s speech during that one session. However, when the clinical screening methods that are currently available can indicate that a patient is affected by dementia, the disease has already impaired cognitive abilities. The best point in time to start a therapy has already passed. Therefore new screening methods are required that can indicate a possible case of dementia much earlier.

We are fortunate to have access to the rich resources of the established Interdisciplinary Longitudinal Study on Adult Development And Aging (ILSE) [17]. In ILSE a wide range of medical parameters as well as biographic interviews have been collected from 1,000 participants over the course of 20 years. Each participant contributed data in up to four measurements. This longitudinal dataset allows us to investigate the predictive capabilities of speech-based dementia screening: we use a participant’s speech at measurement \(t\) and train a classifier to predict the participant’s cognitive diagnosis at measurements 5 years \((t + 5)\) and 12 years \((t + 12)\) in the future.

In Section 2 we describe the selection of data from ILSE we are using for the prediction experiments and in Section 3 we present the baselines we set up to compare the speech-based prediction against. In Sections 4 and 5 we lay out the features we extracted from speech and the experimental results including an analysis of the features that are selected for dementia prediction.

2. DATASET
The speech recordings and cognitive diagnoses that we use in this work are from the Interdisciplinary Longitudinal Study on Adult Development and Aging (ILSE) [17, 18]. Over the course of more than 20 years participants in Germany were invited to take part in four measurements (T1-T4) in which a large corpus of data was collected. T1 was conducted in 1993-1996, T2 in 1997-2000, T3 took part in 2005-2008 and T4 was recorded in 2013-2016. From ILSE’s wealth of data we use two data sources: recordings of biographic interviews and cognitive diagnoses established by psychiatrists using a range of neuropsychological, anamnestic, clinical, and laboratory tests.
The ILSE participants form a group that represents the sampled population (cf. [18, 19]). When the study started, the participants were either 40 or 60 years old. At this age gerontologists expect very few cases of cognitive impairment and indeed most of the ILSE participants had no cognitive impairment when the study began. As participants grew older, some of them developed cognitive impairments: The participants are either diagnosed with aging-associated cognitive decline (AACD) or Alzheimers disease (AD), or are members of the control group. The severity of AACD or AD was not recorded in ILSE. However, some participants dropped out of the study because they felt unable to participate. All participants with very severe AD also dropped out of the study which means we do not have any participants with very severe AD in our dataset.

For the predictive dementia screening we use features extracted from interviews at measurement $t$ to predict the diagnosis at measurements $t+5y$ and $t+12y$. From all the interviews in ILSE we first select those for which a recording of the interview exists and has been manually transcribed. In a second step we select interviews at measurement $t$ so that a diagnosis for the interviewed participant is available at measurements $t+5y$ and/or $t+12y$. As not all participants took part in all measurements there are fewer diagnoses available the longer the time between the measurement from which we use the speech and the measurement for which we predict the diagnosis. Table 1 shows the distribution of diagnoses available for the $t+5y$ and $t+12y$ predictive dementia screening.

<table>
<thead>
<tr>
<th>Prediction</th>
<th>control</th>
<th>AACD</th>
<th>AD</th>
<th># interviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t+5y$</td>
<td>57</td>
<td>14</td>
<td>6</td>
<td>77</td>
</tr>
<tr>
<td>$t+12y$</td>
<td>41</td>
<td>10</td>
<td>6</td>
<td>57</td>
</tr>
</tbody>
</table>

For our feature extraction and experiments we consider two types of transcriptions: manual transcriptions and automatic transcriptions. The manual transcriptions were aligned to the audio using a long audio alignment procedure [18]. We produced the automatic transcriptions using our ILSE speech recognition system [20]. With 58.5% word error rate these automatic transcriptions are highly erroneous, but still enabled them to extract powerful features for dementia state screening [20].

3. DEMENTIA PREDICTION BASELINES

We establish three baselines to evaluate the prediction of dementia based on speech features. These baselines are based on diagnoses and their prevalence which in epidemiology describes the portion of a population that is affected by a disease or condition. We consider the prevalence of AACD and AD in ILSE, i.e. which portion of the ILSE participants is diagnosed with AACD and AD. For their prediction our baselines use (a) only the prevalence, (b) only the diagnosis at measurement $t$, and (c) the prevalence and the diagnosis at measurement $t$.

We use unweighted average recall (UAR) to evaluate our experiments and also these baselines. This metric gives equal weight to all three classes (control, AACD and AD) and is therefore more suitable to this unbalanced dataset than a weighted metric such as accuracy [21]. The classification chance level for a three-class classification is at UAR = 33.3%.

![Baseline Results](image)

**Fig. 1.** Results of the baselines.

3.1. Baseline: prevalence

For this study we selected a set of data from the ILSE corpus according to the criteria in Section 2. However, cognitive diagnoses are available for many more participants in ILSE. In order to get the most robust estimations of the prevalences in the ILSE population, we calculate the prevalences of AD and AACD on the whole set of 2815 ILSE diagnoses. We take into account the two ILSE cohorts (born 1930-3 (C30) and born 1950-52 (C50)) as well as the measurement $t$. For example, for C30 participants the prevalence of AD is 5.8% at T3 and 9.9% at T4. With these prevalences we assign diagnoses for $t+5y$ and $t+12y$ in a Monte Carlo simulation: we randomly assign a diagnosis to each participant according to the distribution provided by the prevalences. After assigning the diagnoses we evaluate the UAR of this assignment. We repeat this for 10,000 iterations. Figure 1 shows the mean and standard deviation of the results of these baseline predictions. Using only the prevalence, the results are hardly better than chance level.

3.2. Baseline: diagnosis at measurement $t$

When nothing but a diagnosis is known, one might assume that the current state is the best predictor for the future. This means using the diagnosis at measurement $t$ as the prediction for measurements $t+5y$ and $t+12y$. Figure 1 shows that the results for these baseline predictions are better than chance level for $t+5y$ but the current diagnosis is not a good predictor for the diagnosis at $t+12y$. 

![Baseline Results](image)
3.3. Baseline: prevalence and diagnosis at measurement $t$

In a second approach to prevalence we take into account the two ILSE cohorts, the measurements $t$ and $t + 5y / t + 12y$, and the diagnosis at measurement $t$. For example, for C30 participants who are healthy controls at T1 the prevalence for T2 is 79% control and 21% AACD. With these prevalences we run a Monte Carlo simulation with 10,000 iterations like in Section 3.1. The results in Figure 1 are better than chance level or using only the prevalence, but are on par with using only the diagnosis at $t$.

4. FEATURES FOR DEMENTIA PREDICTION

We differentiate two categories of features: acoustic features and linguistic features. Linguistic features measure what participants say, while acoustic features measure how they say it. There are three types of acoustic features and twelve types of linguistic features. We have previously successfully leveraged these features for state screening and use them again for predictive screening to maintain comparability. All features are derived from participant speech on a per-interview level which means that each interview is represented by one feature vector.

4.1. Acoustic features

Pause-based Features [22]: The 12 pause-based features include speech pause durations, rates, counts and the ratios between speech pauses and words. These features are calculated from audio, a voice-activity-based speech pause detection and transcriptions.

Speaking Rate Features [22]: We measure speaking rate in words per second and phones per second. These features are extracted from the duration of the audio recording and the transcription. For the calculation of phones per second we also use a pronunciation dictionary.

i-Vector Features [9]: I-Vectors [23, 24] represent speaker characteristics. We extract 128-dimensional i-vectors from the raw audio and regard each dimension as one feature.

4.2. Linguistic features

Lexical Richness [20]: Lexical richness measures the use of vocabulary. We use two such measures: Brunet’s W index [25, 26] and Honoré’s R Statistics [27].

Part-of-Speech (POS) Tags [9]: We use the TreeTagger [28] to automatically extract POS tags and calculate the percentage of each tag. We use a tagset created for written language (POS) [29] and a tagset created specifically for conversational language (conv. POS) [30]. In addition, we group POS tags together into POS categories (e.g. “verbs” or “adjectives”) to produce two more feature types: POS categories and conv. POS categories.

Linguistic Inquiry and Word Count (LIWC) [31]: LIWC uses a dictionary to assign categories to words. The German LIWC dictionary [32] contains 64 categories. We use the percentage of each category as a feature.

Perplexity Features [20]: Perplexity measures how well a statistical language model fits a text. We calculate the perplexity of a speech segment using a model of other segments of that speaker’s speech. The features include the perplexity of 1-gram to 5-gram language models. We extract five different types of perplexity features from five different types of input: one on the textual transcriptions and one each on written POS tags, conversational POS tags, written POS categories and conversational POS categories.

Between-Speaker Perplexity [20]: In addition to the within-speaker perplexity described above, we also extract the between-speaker perplexity. We calculate this type of perplexity for text only. For this perplexity, language models are built on all but one speaker’s speech and then evaluated on that speaker’s speech.

5. EXPERIMENT & RESULTS

For the first time we use the features extracted from speech at measurement $t$ to predict the diagnosis at measurement $t + 5y$ and $t + 12y$. These experiments are three-class classification for the cognitive diagnoses control, AACD and AD. The predictive classifiers use only the speech-based features from measurement $t$ and do not have any information about the participants’ cognitive diagnoses at measurement $t$ for their prediction.

We extract one feature vector per recording, and train and evaluate classifiers in a leave-one-person-out cross-validation. As in our previous work [20] a Gaussian classifier achieved the best classification results which is why we are only using diagonal Gaussian classifiers. Our experiments are based on the implementations in scikit-learn [33].

5.1. The first predictive dementia screening from speech

This experiment is the first predictive dementia screening from speech. For the predictions we use a feature set that we previously optimized for state screening [9]. Figure 2 shows the result of this dementia prediction for $t + 5y$ and $t + 12y$, and the result that we obtained for state screening [9] as a comparison. The prediction for $t + 5y$ achieves good results at 57.3% UAR with features from manual transcriptions and 60.2% with features from automatic transcriptions. These results are well above the baselines and there is a significant difference in classification results ($p < 0.005$) compared to state screening.
According to a \( \chi^2 \) test with three degrees of freedom. The prediction for \( t + 12y \) produces results that are comparable to the baselines. The result for features from manual transcriptions at 38.7% UAR is just above these baselines while the performance for features from automatic transcriptions is below the baselines at 32.9% UAR. The classification results differ significantly (\( p < 0.001 \)) from the state screening results.

This feature set is optimized for state screening. We observe that the performance gradually declines as the prediction moves further into the future from \( t \) to \( t + 5y \) and finally \( t + 12y \). However, the prediction for \( t + 5y \) works well even with the feature set optimized for state screening. The significant differences between state screening and predictive screening mean that the predictive classifiers learn to predict diagnoses. In Section 5.2 we investigate whether we can optimize the feature sets to produce prediction results comparable to state screening results.

Following this procedure we extract feature sets from 54 folds for \( t + 5y \). Figure 3 shows the ranking of how often each feature was selected for the \( t + 5y \) prediction. Out of the 444 features only 96 feature were selected in at least one fold. No features were selected in all folds, but the most commonly selected feature (the “sex” category from LIWC) was selected in 52 folds. 22 features were only selected in one fold. Analyzing the ranking in Figure 3 we see that mostly linguistic features were selected. The most commonly selected feature types are LIWC, conversational POS, i-Vector and written POS. The three most common features are all LIWC features: “sex”, “nonfluency” and “down”.

For the \( t + 12y \) prediction we extract feature sets from 50 folds. Figure 4 shows the ranking of the features selected for the \( t + 12y \) prediction. 136 out of the 444 features were selected at least in one fold. Again, no feature was selected in all folds, but the most common one appears in 48 folds. 36 out of the 136 features were selected in only one fold. We observe that linguistic features are again the most commonly selected features and acoustic features were selected even less often for \( t + 12y \) than for \( t + 5y \). The most common features types are again LIWC, conversational POS and written POS. The three top-ranking features are again LIWC type features: “anger”, “sad” and “negative emotion”.

We analyze these two feature rankings and compare them to the ranking that we obtained for state screening [9]. In this way we achieve some insight into the dementia prediction process. In all three cases i-vector, LIWC, conversational POS and written POS were the most important features. However, the further the prediction moves into the future from \( t \) to \( t + 5y \) and finally \( t + 12y \), the less often acoustic features are selected. While for state screening the top-ranked feature is an acoustic feature, for the \( t + 5y \) prediction the 10th-ranked feature is an acoustic feature and for the \( t + 12y \) prediction the highest ranked acoustic feature has rank 68 and was only selected six times. It appears that the features do not extract information that allows conclusions about the future from the acoustics, but that the linguistics contain information useful for dementia prediction.

### 5.2. Optimizing feature sets for dementia prediction

#### 5.2.1. Selecting features for dementia prediction

The features described in Section 4 amount to a 444-dimensional feature vector. Using all these features the feature dimensionality would be larger than the number of training samples. This therefore presents itself as an under-determined machine learning problem. We therefore seek to reduce the dimensionality by using a nested forward feature selection to select features: We run leave-one-person-out cross-validation to produce folds of training sets and test sets. On the training set of each fold we run a forward feature selection in a second level of leave-one-person-out cross-validation. From each of these runs of forward feature selection we extract the set of features that produced the best result. In total, we obtain one set of features for each fold of the leave-one-person-out cross-validation.

### 5.2.2. Predicting dementia with selected features

With the rankings from the nested forward feature selection (Figures 3 and 4) we produce feature sets for dementia prediction. We create feature sets by adding features in order of their appearance in the ranking like we did for state screening [9]. Going through the ranking we create a new feature set by using the previous feature set and adding the next features that have been selected the same number of times: In the case of the \( t + 5y \) prediction (compare Figure 3) we create the first feature set using only the top-ranking feature since there are no other features that were selected 52 times. The next two features both appear in 51 folds, so they are added to the previous feature set so that the second feature set contains the top...
three features. Continuing this procedure the third feature set contains five features and the fourth feature set contains nine features. We create feature sets until the features that appear only once have been included in the last feature set.

With these feature sets we run dementia prediction experiments as before (Section 5.1). For \( t + 5y \) the best achieved performance is 73.3% UAR using manual transcripts and 81.2% using automatic transcripts. This result is obtained with a 53-dimensional feature set that contains mainly linguistic features, but also acoustic features: the two i-Vector features ranked 10\(^{th}\) and 15\(^{th}\). Although an improvement of 16.0% absolute over the result in Section 5.1 the \( \chi^2 \) test with three degrees of freedom does not show a significant difference. Figure 5 shows the confusion matrix for this result using the manual transcripts. No person that will develop AD five years after the recording is predicted to stay cognitively healthy and only one is mistakenly predicted to develop AACD. For AACD we get a similar picture: Only one person was incorrectly predicted to stay healthy and only two persons were incorrectly predicted to develop AD. For the people who stayed cognitively healthy, we observe that nearly half of them are predicted to develop cognitive impairment. If the classifier has to make errors, this is the direction we would like the errors to be: for the affected individual it is better to undergo preventive therapy that is not
strictly needed than to miss out on urgently needed preventive measures.

For the \( t + 12y \) prediction the best classifier achieves a performance of 75.7% UAR using features extracted from manual transcriptions. This is obtained with both 19 and 21 features, all of which are linguistic features since for \( t + 12y \) the acoustic features are ranked much lower. Using automatic transcriptions the result is only at baseline level, the reasons for this are yet to be investigated. The result using manual transcriptions is an 35.0% absolute improvement over the results obtained in our first dementia prediction experiment (Section 5.1). There is a significant difference \((p < 0.005)\) between the two results according to a the \( \chi^2 \) test with three degrees of freedom. Figure 6 shows the confusion matrix of this result using manual transcriptions. The qualitative results are similar to the results for the \( t + 5y \) prediction. No person who will develop a cognitive impairment in 12 years’ time has been classified as cognitively healthy. One person who will develop AD has been predicted to develop AADC, and one person who will develop AADC has been classified to develop AD. However, again the classifier makes more mistakes for persons who will stay cognitively healthy.

We have obtained results that clearly beat the baselines in Section 3. Optimizing the feature sets in this manner produced good improvements over the results obtained from the feature set optimized for state screening. We observe that acoustic features are less often used the further one predicts into the future. The linguistic feature seem to be more suitable to capture development that is not yet noticeable by other means. It is therefore necessary to view each of these tasks individually and accordingly optimize feature sets individually. The results with the optimized feature sets are not too far away from the results that we observed for state screening (Figure 2). This is a very promising result considering that predictive screening allows for a hint into future development.

6. CONCLUSIONS

We have presented predictive screening for dementia from speech for the first time. Using speech extracted from biographic interviews we predicted the cognitive diagnosis 5 and 12 years later. For the interpretation of results for this ambitious task we created baselines based on the prevalence of the cognitive diagnoses in ILSE as well as the participant’s prior diagnoses. In contrast to the baselines the speech-based prediction uses neither prevalence nor existing diagnoses. Using the best performing feature set from our dementia state screening we see that the classifiers learn to predict future diagnoses. For the prediction for \( t + 5y \) it produces a good performance while for the \( t + 12y \) predictions the results are at baseline level. Selecting feature sets in a nested forward selection specifically for the predictions improves on these results. The best result for the \( t + 5y \) prediction is 73.3% UAR, an improvement of 16.0% absolute over the over the result with the non-optimized feature set. The best result for \( t + 12y \) prediction is 75.7% UAR which is a significant 35.0% improvement over the non-optimized result. We observe, that the further predictions are made into the future, the less often acoustic features are selected by the feature selection. In both cases the classification results differ significantly from the baselines for the two non-control classes which lets us conclude that a predictive screening for dementia from speech is possible.
7. REFERENCES


