Selecting Features for Automatic Screening for Dementia based on Speech

Jochen Weiner and Tanja Schultz

Cognitive Systems Lab, University of Bremen, Germany
jochen.weiner@uni-bremen.de

Abstract. As the population in developed countries ages, larger numbers of people are at risk of developing dementia. In the near future large-scale time- and cost-efficient screening methods will be needed. Speech can be recorded and analyzed in this manner, and as speech and language are affected early on in the course of dementia, automatic speech processing can provide valuable support for such screening methods.
We have developed acoustic and linguistic features for dementia screening and established that a combination of acoustic and linguistic features provides the best results. However, our full set of 429 fine-grained features from 15 feature types is too large to train a robust model on limited training data. We therefore need to select features to use for dementia screening. We employ forward feature selection nested in a cross-validation and identify the most commonly selected features. Both acoustic and linguistic features from seven different feature types are selected. Using sets of these features we obtain a 0.819 unweighted average recall which is a strong improvement over previous results.

Keywords: computational paralinguistics · dementia screening · feature selection

1 Introduction

The demographic development in Germany and other countries is accompanied by a severe increase in geriatric diseases. Their most common representative is dementia, a chronic progressive disease that is accompanied by loss of autonomy in everyday life. As no curative therapy is known [6], early secondary prevention measures are of great importance. Current diagnostic procedures require a thorough examination by medical specialists, which unfortunately are too cost- and time-consuming to be provided frequently on a large scale. Since speech and language capacity is a well established early indicator of cognitive deficits including dementia [1, 4], speech processing methods offer great potential to fully automatically screen for prototypic indicators in real-time and to present analyses and results such that medical specialists can include them as an additional information source when diagnosing cognitive deficits. Acoustic features [20, 7, 28, 12, 33] and linguistic features [8, 18, 13, 12, 9, 30, 36, 27, 11] have been used in a classification task that aims to distinguish speakers affected by dementia from cognitively health speakers using just their speech.

We are fortunate to have access to the rich resource of conversational speech data from the established *Interdisciplinary Longitudinal Study on Adult Development And Aging* (ILSE) [21] in which a range of medical parameters and more than 10,000 hours of interviews were recorded from more than 1,000 subjects over the course of 20 years. We established first results on dementia screening from ILSE interviews based on both acoustic [33] and linguistic features [31] and found that the combination of acoustic and linguistic features gives best results. With over 400 features from three acoustic and twelve linguistic feature types our full set of features is too large to train the robust classifier we need for automatic screening. Therefore we need to select the features which provide the best information for the classifier to exploit. All our features have been hand-crafted to capture aspects of the changes in speech and language characteristics caused by dementia, are automatically extractable from speech and have contributed well to dementia screening experiments [33, 31]. We use a data-driven approach to select the features which the classifier will use. In the past we have employed mutual information and ANOVA with unsatisfying results. In this paper we use a forward feature selection nested in a cross-validation and identify the features which are selected in the most cross-validation folds. We employ the sets of the most commonly selected features for dementia screening and achieve a strong enhancement in performance over our previous results [33, 31].

2 Database

The ILSE study acquired massive amounts of data for research in participants' personality, cognitive functioning, subjective well-being and health. Over the course of more than 20 years participants contributed to four measurements. In each measurement participants took part in a range of medical, psychological, cognitive, physical, and dental examinations, as well as semi-standardized biographic interviews. From this wealth of data we use the participant's speech recorded in biographic interviews, and their cognitive diagnoses. The participants are either diagnosed as cognitively healthy (control), with aging-associated cognitive decline (AACD) or Alzheimer's disease (AD).

In this paper we use 98 interviews from 62 participants, for which we have manual transcriptions with speaker turn annotations (no time alignments) plus the cognitive diagnoses of the participants [33]. The ILSE participants form a group that represents the sampled population (cf. [32, 14]). When the study started, the participants were at an age which is considered to be young in gerontological terms and thus most of the ILSE participants had no cognitive deficits when the study began. As the study progressed, some of the participants developed cognitive deficits as anticipated by the prevalence of cognitive impairment with age [17, p. 20]. In our dataset there are 80 interviews with healthy controls, 13 with AACD participants and 5 interviews with participants with AD.

The interviews were recorded with only one microphone, i.e. their speech occurs on the same audio channel. Since dementia screening focuses on the participant, we first select the participant's speech segments from the interview recordings. The manual transcriptions provide the order of speaker contributions but no time-alignment. Therefore, we perform long audio alignment [32] to infer speaker segmentation and subsequently select 230 hours of speech.

3 Features for Dementia Screening

We differentiate two categories of features: acoustic features and linguistic features. Linguistic features measure *what* participants say, while acoustic features measure *how* they speak. All features are derived from participant speech on a per-interview level.

3.1 Acoustic Features

Acoustic features measure how participants speak. They are extracted from voice-activity labels, transcriptions, and the raw audio. We have three types of acoustic features:

Pause-based Features [33]: A very pronounced difference between people with dementia and healthy controls that has been reported is that people afflicted with dementia tend to hesitate more often and make longer pauses [12]. The 12 speech pause-based features include speech pause durations, rates, counts and the ratios between speech pauses and words (for a full description see Weiner et al. [33]). These features are calculated from audio, speech pauses indicated by our voice activity detection system [33] and transcriptions.

Speaking rate [33]: Longer pauses and more hesitations in speech from people with dementia imply that more time is needed to convey words. Speaking rate is measured in words per second and phones per second. Audio, transcriptions and a pronunciation dictionary are used to calculate these features.

i-Vector Features: An i-vector [5, 25] is a compact representation of an individual speaker's acoustic characteristics. In speaker recognition and diarization they are used to distinguish between speakers. For the task of dementia screening we use i-vectors to distinguish between cognitive diagnoses, not between speakers. We use the fact that dementia affects a speaker's speech characteristics. Since the i-vector represents these characteristics we expect to see dementia-specific differences between i-vectors extracted from speech of healthy participants and those extracted from speech of participants with dementia. I-vectors are extracted from the raw audio. For dementia screening we regard each dimension of the 128-dimensional i-vector as one feature.

3.2 Linguistic Features

We use linguistic features to measure changes to vocabulary, word usage and sentence structure that are caused by dementia. Our linguistic features operate at the word surface level of the transcriptions. There are twelve types of linguistic features:

Lexical Richness [31]: Lexical richness measures the participants' use of their vocabulary. Changes to spoken language characteristics caused by dementia alter which words speakers can access and how they use these words. This is visible on the word surface level and can be measured by observing changes in lexical richness. We use two measures for the lexical richness of the transcriptions: Brunet's W index [3, 29] and Honoré's R Statistics [10].

Linguistic Inquiry and Word Count (LIWC) [15]: Developed as a tool in psychological research, LIWC has been shown to produce markers for a variety of individual differences including cognitive processing [16, 26]. Dementia, as one form of cognitive impairment, highly affects cognitive processes that can be investigated using LIWC. For this reason, LIWC has already been used successfully in dementia screening [2, 19, 11]. LIWC uses a dictionary to assign each word to a category and calculates the percentage of each LIWC category in a transcription. The 64 categories in the German dictionary [35] cover basic linguistic dimensions, psychological processes, relativity and personal concerns. We use the percentage of each category as one feature.

Part-of-Speech (POS) Tags: Words with similar grammatical properties can be grouped together by POS tags. Each tag represents the grammatical role a word can take in a sentence and thus POS tags can be used to indicate grammatical properties of participants' speech. We employ the TreeTagger [23] to automatically extract POS tags and calculate the percentage of each tag. We use both the traditional tagset created for written language [22] and a tagset created specifically for conversational language [34]. Furthermore, we group tags together to form POS categories. Thus we have four types of POS features: written POS (51 features), written POS categories (11 features), conversational POS (57 features), and conversational POS categories (12 features).

Perplexity Features [31]: Perplexity measures how well a statistical language model fits a text. The lower the perplexity the better the model is able to predict the text. If people with dementia use simpler sentences and repeat themselves more often, their speech is more predictable than the speech of healthy speakers. Thus the speech of a person with dementia will have a lower perplexity. We calculate two different types of perplexity features, each comprising 15 features:

- *Within-Speaker Perplexity Features* measure the predictability of speakers' speech relative to another segment of their own speech. We extract five sets of these features: from text, from written and conversational POS tags, and from written and conversational POS tag categories.
- *Between-Speaker Perplexity Features* are obtained as a measure of the predictability of a speaker's speech in comparison to other speakers. We extract these features only from text.

4 Feature Selection

The features described in Section 3 amount to a 426-dimensional feature vector. As this feature dimension is much higher than our sample size of 98 interviews we cannot expect to train a robust classifier using the whole set of features. In the past we have employed mutual information and ANOVA with unsatisfying results. In this paper we therefore employ forward feature selection to select the features which we will use for dementia screening.

For dementia screening we train and evaluate classifiers in a leave-one-person-out cross-validation. Each participant contributed to ILSE in more than one measurement, so the model is trained on the data from all but one participant and then evaluated on the participant that was not used in training. This cross validation ensures that a participant is never in both the training set and the test set at the same time. We use *unweighted average recall (UAR)* to evaluate our experiments. This metric gives equal weight to all three classes (control, AACD and AD). Since the distribution of diagnoses in ILSE is determined by their natural occurrence, UAR is more suitable than a weighted metric such as accuracy [24]. The chance level for a three-class classification is at UAR = $1/3$.

On the training set of each cross-validation fold, we run a forward feature selection in a second level of leave-one-person-out cross-validation using a diagonal Gaussian classifier. For each of the 62 folds we extract the set of features that produced the best result. The median set contains 26 features. Overall, 88 of our total 426 features have contributed to the selected feature set in at least one fold. Figure 1 shows how often each of these features was selected and ranks the features by the number of times they were selected. The most common feature was selected in 61 of the folds while 14 features were selected only once. Both acoustic and linguistic features were selected, but the majority of the selected features are linguistic features. Selected acoustic features are from the i-vector and pause-based feature types. Linguistic features were selected from the LIWC, conversational POS, conversational POS categories, written POS and written POS categories feature types. No features were selected from the speaking rate or the six perplexity feature types.

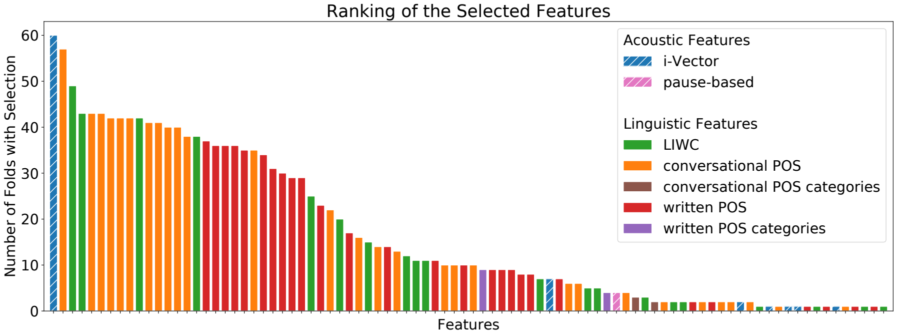

Fig. 1. Ranking of the 88 selected features by the number of times they were selected. The legend points out the acoustic and linguistic feature types.

The three most commonly selected features were selected in more than three quarters of the folds. In Figure 2 we show the value ranges of these features for the three different diagnoses (control, AACD and AD). The figure shows a clear difference between the medians of the features for the different diagnoses. We also observe that the variance of the features is greatly reduced for people with AD. Overall, the figure shows that the three most commonly selected features allow for differentiation between the diagnoses and are therefore good selections as features for dementia screening.

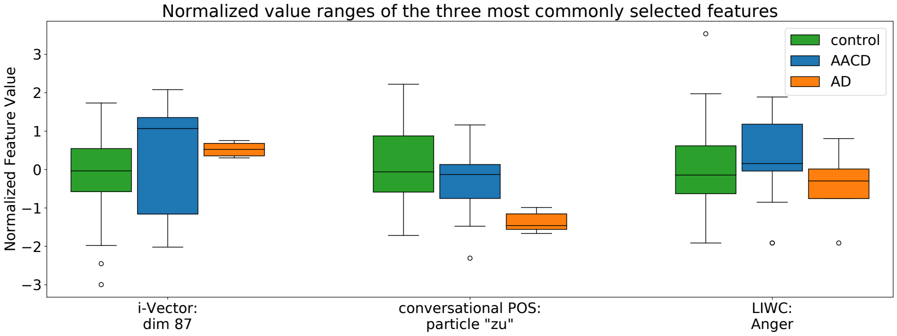

Fig. 2. Value ranges of the three most commonly selected features. For the visualization we applied a z-normalization to the features.

5 Dementia Screening using Selected Features

From the results of the feature selection (Section 4) we infer feature sets for our dementia screening experiments. We create feature sets by including features in the order of their ranking in Figure 1: The first feature set contains only the one most commonly selected feature. The second feature set contains the first and second ranking features, the third set the first, second and third features. The features ranking four through six were each selected the same number of times. Therefore we include these in the next feature set together so that the fourth feature set contains the features ranking one through six. In this manner we create 36 feature sets, where each set is a complete subset of all the larger feature sets.

Using each of these feature sets, we perform a dementia screening experiment as described above, using a leave-one-person-out cross-validation and a Gaussian classifier. Again we use UAR to evaluate the experiments. The results of these experiments are shown in Figure 3. The feature sets that we have obtained from the forward feature selection clearly outperform feature sets of the same size obtained by mutual information or ANOVA feature selection [33, 31]. We see

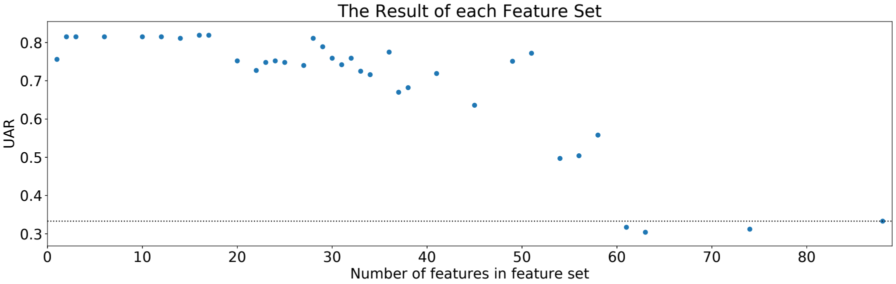

Fig. 3. Results using the feature sets inferred from the forward feature selection. The dotted line represents chance level at UAR = 1/3

very good results for one feature at 0.756 UAR and for sets of two to twelve features with equal results at 0.815 UAR. There is a further slight improvement in performance: Using feature sets of 16 and 17 features, the experiment achieves its maximal UAR of 0.819. These feature sets contain the sixteen and seventeen most commonly selected features, respectively. In the 16-feature-set each feature was selected at least 37 times, the 17-feature-set includes one more feature which was selected 36 times. Both feature sets contain i-vector, conversational POS, LIWC and written POS type features. As the feature set dimensionality increases further, the results deteriorate dramatically to results at or below chance level when the feature set includes features that the forward selection selected in less than five folds.

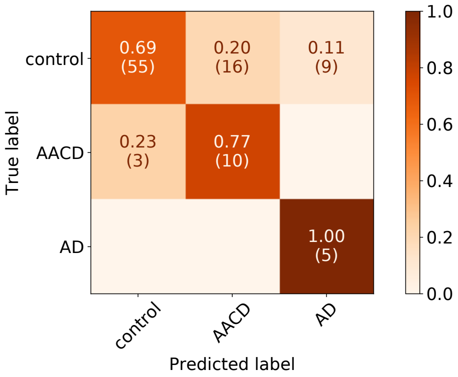

Fig. 4. Confusion matrix for the result of 0.819 UAR using 16 or 17 features.

Figure 4 shows a confusion matrix for this best result of UAR 0.819, which is the same for both the 16- and 17-feature set. The confusion matrix shows the enhanced performance over our previous results [33, 31] very clearly. The confusions between all classes have been reduced, all participants with AD have been classified correctly and there are no more confusions between participants with AACD and participants with AD.

6 Conclusions

The increasing numbers of elderly people who need to be screened for dementia require automatic highly accurate approaches to support clinicians in making dementia diagnoses. We developed a fine-grained set of 429 features of three acoustic and twelve linguistic feature types. Using this high number of features with our set of 98 interviews we cannot expect to train a robust classifier. Therefore we need to select which features to use in dementia screening. Investigating the results of a forward feature selection we show that both acoustic and linguistic features are included in the best feature sets and that the value ranges of the most commonly selected features indicate that these are good selections. Inferring feature sets of the most commonly selected features, we obtain the best result at 0.819 UAR with 16- and 17-dimensional feature sets. We also get very good results with two (one acoustic and one linguistic) features. This result far outperforms our previous results and is an important step towards highly accurate screening for dementia based on speech.

7 Acknowledgements

We thank Johannes Schröder at the University of Heidelberg and the University Hospital Heidelberg, Germany for providing us with the ILSE data and for sharing with us his expertise on speech and language use of people with dementia.

References

1. Appell, J., Kertesz, A., Fisman, M.: A study of language functioning in Alzheimer patients. Brain and language **17**(1), 73–91 (1982)
2. Asgari, M., Kaye, J., Dodge, H.: Predicting mild cognitive impairment from spontaneous spoken utterances. Alzheimer's & Dementia: Translational Research & Clinical Interventions **3**(2), 219 – 228 (2017)
3. Brunet, E.: Le vocabulaire de Jean Giraudoux, structure et évolution. Slatkine, Geneva (1978)
4. Bucks, R., Singh, S., Cuerden, J.M., Wilcock, G.K.: Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. Aphasiology **14**(1), 71–91 (2000)
5. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing **19**(4), 788–798 (2011)
6. Deutsche Gesellschaft für Psychiatrie und Psychotherapie, Psychosomatik und Nervenheilkunde (DGPPN), Deutsche Gesellschaft für Neurologie (DGN): S3-Leitlinie "Demenzen". https://www.dgppn.de/_Resources/Persistent/ade50e44afc7eb8024e7f65ed3f44e995583c3a0/S3-LL-Demenzen-240116.pdf (2016), accessed 16.04.2018

7. Espinoza-Cuadros, F., Garcia-Zamora, M.A., Torres-Boza, D., Ferrer-Riesgo, C.A., Montero-Benavides, A., Gonzalez-Moreira, E., Hernandez-Gómez, L.A.: 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)
8. Hakkani-Tür, D., Vergyri, D., Tür, G.: Speech-based automated cognitive status assessment. In: INTERSPEECH 2010 – 11th Annual Conference of the International Speech Communication Association. pp. 258–261 (2010)
9. Hernández-Domínguez, L., García-Cano, E., Ratté, S., Sierra-Martínez, G.: Detection of Alzheimer's disease based on automatic analysis of common objects descriptions. In: Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning. pp. 10–15 (2016)
10. Honoré, A.: Some Simple Measures of Richness of Vocabulary. Association for Literary and Linguistic Computing Bulletin **7**(2), 172–177 (1979)
11. Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M.L., Ogar, J.: Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. pp. 27–37 (2014)
12. Khodabakhsh, A., Yesil, F., Guner, E., Demiroglu, C.: Evaluation of linguistic and prosodic features for detection of Alzheimer's disease in Turkish conversational speech. EURASIP Journal on Audio, Speech, and Music Processing **2015**(1), 1–15 (2015)
13. Lehr, M., Prud'hommeaux, E.T., Shafran, I., Roark, B.: 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)
14. Martin, P., Martin, M.: Design und Methodik der Interdisziplinären Längsschnittstudie des Erwachsenenalters. In: Martin, P., Ettrich, K.U., Lehr, U., Roether, D., Martin, M., Fischer-Cyrulies, A. (eds.) Aspekte der Entwicklung im mittleren und höheren Lebensalter: Ergebnisse der Interdisziplinären Längsschnittstudie des Erwachsenenalters (ILSE), pp. 17–27. Steinkopff (2000)
15. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: LIWC 2001. Erlbaum (2001)
16. Pennebaker, J.W., Graybeal, A.: Patterns of natural language use: Disclosure, personality, and social integration. Current Directions in Psychological Science **10**(3), 90–93 (2001)
17. Prince, M., Wimo, A., Guerchet, M., Ali, G.C., Wu, Y.T., Prina, M.: World Alzheimer Report 2015. The Global Impact of Dementia: an Analysis of Prevalence, Incidence, Cost and Trends. Alzheimer's Disease International, London (2015)
18. Prud'hommeaux, E.T., Roark, B.: Extraction of narrative recall patterns for neuropsychological assessment. In: INTERSPEECH 2011 – 12th Annual Conference of the International Speech Communication Association. pp. 3021–3024 (2011)
19. Sadeghian, R., Schaffer, J.D., Zahorian, S.A.: Speech Processing Approach for Diagnosing Dementia in an Early Stage. In: INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association. pp. 2705–2709 (2017)
20. Satt, A., Hoory, R., König, A., Aalten, P., Robert, P.H.: 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)

21. Sattler, C., Wahl, H.W., Schröder, J., Kruse, A., Schönknecht, P., Kunzmann, U., Braun, T., Degen, C., Nitschke, I., Rahmlow, W., Rammelsberg, P., Siebert, J.S., Tauber, B., Wendelstein, B., Zenthöfer, A.: Interdisciplinary Longitudinal Study on Adult Development and Aging (ILSE), pp. 1213–1222. Springer, Singapore (2017)
22. Schiller, A., Teufel, S., Stöckert, C.: Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset) (1999)
23. Schmid, H.: Improvements in Part-of-Speech Tagging with an Application to German. In: Proceedings of the ACL SIGDAT-Workshop. pp. 47–50 (1995)
24. Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J.K., Baird, A., Elkins, A., Zhang, Y., Coutinho, E., Evanini, K.: The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language. In: INTERSPEECH 2016 – 17^{th} Annual Conference of the International Speech Communication Association. pp. 2001–2005 (2016)
25. Shum, S.H., Dehak, N., Dehak, R., Glass, J.R.: Unsupervised methods for speaker diarization: An integrated and iterative approach. IEEE Transactions on Audio, Speech, and Language Processing **21**(10), 2015–2028 (2013)
26. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology **29**(1), 24–54 (2009)
27. Thomas, C., Kešelj, V., Cercone, N., Rockwood, K., Asp, E.: Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech. In: IEEE International Conference Mechatronics and Automation. vol. 3, pp. 1569–1574 Vol. 3 (2005)
28. Tóth, L., Gosztolya, G., Vincze, V., Hoffmann, I., Szatlóczki, G.: Automatic detection of mild cognitive impairment from spontaneous speech using ASR. In: INTERSPEECH 2015 – 16^{th} Annual Conference of the International Speech Communication Association. pp. 2694–2698 (2015)
29. Tweedie, F.J., Baayen, R.H.: How Variable May a Constant be? Measures of Lexical Richness in Perspective. Computers and the Humanities **32**(5), 323–352 (1998)
30. Wankerl, S., Nöth, E., Evert, S.: An Analysis of Perplexity to Reveal the Effects of Alzheimer's Disease on Language. In: 12th ITG Conference on Speech Communication. pp. 254–258 (2016)
31. Weiner, J., Engelbart, M., Schultz, T.: Manual and Automatic Transcription in Dementia Detection from Speech. In: INTERSPEECH 2017 – 18^{th} Annual Conference of the International Speech Communication Association. pp. 3117–3121 (2017)
32. Weiner, J., Frankenberg, C., Telaar, D., Wendelstein, B., Schröder, J., Schultz, T.: 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). pp. 718–725 (2016)
33. Weiner, J., Herff, C., Schultz, T.: Speech-Based Detection of Alzheimer's Disease in Conversational German. In: INTERSPEECH 2016 – 17^{th} Annual Conference of the International Speech Communication Association. pp. 1938–1942 (2016)
34. Westpfahl, S., Schmidt, T.: Folk-gold – a gold standard for part-of-speech tagging of spoken german. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 1493–1499 (2016)
35. Wolf, M., Horn, A.B., Mehl, M.R., Haug, S., Pennebaker, J.W., Kordy, H.: Computergestützte quantitative Textanalyse. Diagnostica **54**(2), 85–98 (2008)
36. Zhou, L., Fraser, K.C., Rudzicz, F.: Speech recognition in alzheimer's disease and in its assessment. In: INTERSPEECH 2016 – 17^{th} Annual Conference of the International Speech Communication Association. pp. 1948–1952 (2016)