Interspeech 2018
2-6 September 2018, Hyderabad

Investigating the Effect of Audio Duration on Dementia Detection using Acoustic Features

Jochen Weiner[1], *Miguel Angrick*[1], *Srinivasan Umesh*[2], *Tanja Schultz*[1]

[1]Cognitive Systems Lab, University of Bremen, Germany
[2]Department of Electrical Engineering, Indian Institute of Technology (IIT) Madras, India

jochen.weiner@uni-bremen.de

Abstract

This paper presents recent progress toward our goal to enable area-wide pre-screening methods for the early detection of dementia based on automatically processing conversational speech of a representative group of more than 200 subjects. We focus on conversational speech since it is the natural form of communication that can be recorded unobtrusively, without adding stress to subjects, and without the need of controlled clinical settings. We describe our unsupervised process chain consisting of voice activity detection and speaker diarization followed by extraction of features and detection of early signs of dementia. The unsupervised system achieves up to 0.645 unweighted average recall (UAR) and compares favorably to a system that was carefully designed on manually annotated data. To further lower the burden for subjects, we investigate UAR over speech duration, and find that about 12 minutes of interview are sufficient to achieve the best UAR.

Index Terms: computational paralinguistics, dementia detection, representative dataset, unsupervised diarization

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, 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 on a large scale. Since speech and language capacity is a well established early indicator of cognitive deficits including dementia [1, 2], 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.

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) [3] 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 detection from ILSE interviews based on both acoustic [4] and linguistic features [5] and found that the combination of acoustic and linguistic features gives best results. Furthermore, we showed that linguistic features derived from automatic speech recognition (ASR) output perform as well as those derived from manual transcripts [5].

In this paper we aim for two goals: first, to investigate the potential of reliable large-scale dementia screening by fully automatic unsupervised speech processing methods and, second, to lower the burden and screening time for both subjects and medical personnel by investigating the minimal amount of speech data required for dementia detection from conversational speech. Regarding fully automatic processing, we no longer rely on any manual transcription, manual speaker segmentation or any other knowledge that requires manual annotation. Rather, we develop and apply fully unsupervised speaker diarization followed by speech recognition to generate transcriptions for a larger set of interviews. This enables us to extend the amount of processed data to 241 interviews covering about 550 hours of speech from a representative subset of 218 ILSE participants. Using this data we first automatically identify participants' speech and then screen for cognitive deficits. To the best of our knowledge, published prior work on dementia detection has so far relied on speaker monologues [6, 7, 8, 9, 10] or on manually identified speaker segments [11, 12, 13, 14, 15, 16] recorded in controlled settings from a rather small number of speakers. The results on dementia detection reported in this paper are thus the first that are based on conversational speech processed by speaker diarization to identify speaker segments in an unsupervised fashion. While the current analysis is limited to acoustic features for dementia detection, we plan future experiments and expect based on our prior results that a combination with linguistic features derived from ASR output will further improve the detection. Furthermore, we investigate the minimal amount of speech data that is required to reliably detect signs of dementia from spoken language. For this purpose we randomly sub-sample the interviews data to speech segments as short as 2.5 minutes and compare detection performance with longer segments.

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 Alzheimers disease (AD). The severity of AACD or AD was not documented in ILSE. We do, however, know that some participants dropped out of the study because they felt unable to participate. It is very likely that participants with very severe AD are among those who dropped out and we assume that there are no recorded participants with very severe AD. While interviews in the first

10.21437/Interspeech.2018-57

measurement last up to 10 hours, they were shorter as the study progressed. However, the duration of interviews were not substantially different for different diagnoses.

The ILSE participants form a group that represents the sampled population (cf. [17, 18]). When the study started, the participants were either 40 or 60 years old. According to gerontological terms, people of this age are considered young and thus cognitive impairment is expected to be rare. 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 [19, p. 20].

At this time, there are 98 interviews from 74 participants, for which we have manual transcriptions with speaker turn annotations (no time alignments) plus the cognitive diagnoses of the participants [4]. We refer to this data as *transcribed dataset*. In addition, we have 241 interviews conducted with 218 participants along with cognitive diagnoses in all four measurements, for which neither transcriptions nor turn and speaker segmentation is available. We refer to these interviews as *untranscribed data set*. The distribution of cognitive diagnoses in these two data sets is summarized in Table 1.

Table 1: *Cognitive diagnoses in the transcribed (*T*) and untranscribed (*U*) data set over four measurements.*

	control		AACD		AD		Total	
Measurement	**T**	**U**	**T**	**U**	**T**	**U**	**T**	**U**
1 (1993-1996)	51	113	4	17	-	-	55	130
2 (1997-2000)	19	67	8	22	-	-	27	89
3 (2005-2008)	10	8	1	2	5	11	16	21
4 (2013-2016)	-	-	-	-	-	1	-	1
Total	80	188	13	41	5	12	98	241

3. Data Preparation

The interviews were recorded with only one microphone, i.e. their speech occurs on the same audio channel. Since dementia detection focuses on the participant, we first select the participant's speech segments from the interview recordings.

The transcribed data set provides the order of speaker contributions but no time-alignment. Therefore, we perform long audio alignment [17] to infer speaker segmentation and subsequently select in total 230 hours of speech from 74 participants.

From the untranscribed data set we fully automatically select participant speech in three subsequent steps: voice activity detection (VAD), speaker diarization, and assignment of diarization clusters to participants and interviewers. In total we selected 550 hours of trustworthy speech from 218 participants.

3.1. Voice Activity Detection

The voice activity detection (VAD) system is a Hidden-Markov-Model recognizer with two Gaussian Mixture Models (GMM): one for speech and one for non-speech. The GMMs have 128 Gaussians each and are trained using BioKIT [20]. As features we use Mel-frequency cepstral coefficients with first and second order derivatives plus zero crossing rate. The models are trained on a small set of 12 interviews (15 hours) for which we have manual transcriptions with a sentence-level time-alignment.

We run a first VAD decoding pass on the interviews with these models, then adapt the models to each audio recording using Maximum Likelihood Linear Regression (MLLR). In a second decoding pass the transformed models are used to label speech and non-speech segments. By a final post-processing pass we ensure that non-speech segments are at least 0.2 s long and speech segments are at least 0.1 s long.

We evaluated the VAD system on held-out data (6 interviews, 9.5 hours) for which sentence-level time-aligned transcriptions are given and found that the VAD marks beginning and end of speech segments very precisely. Since the VAD results also provide labels for short pauses, they are superior to sentence-level time-aligned manual transcriptions.

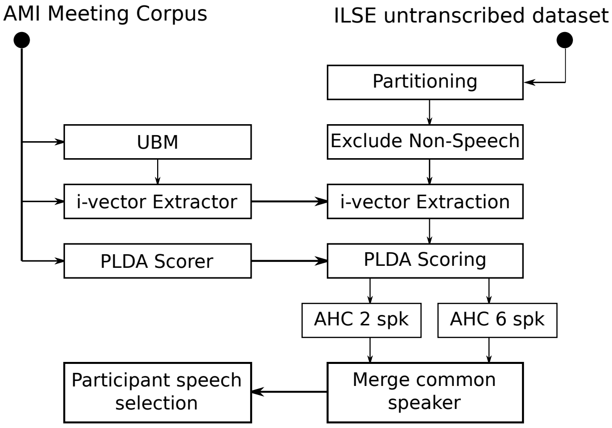

Figure 1: *The diarization system used to extract trustworthy participant speech data.*

3.2. Speaker Diarization

Training a diarization system requires audio recordings with speaker segmentation. However, ILSE only has 15 hours of interviews with speaker segmentation available for training which is too little to build a reliable diarization model. For this reason we use the AMI Meeting Corpus [21] as training data. AMI contains 100 hours of speech from meetings and is thus comparable to the ILSE interview setting. Figure 1 gives an overview of the complete diarization system and the use of data sets. No ILSE data is used in training the diarization system which therefore performs unsupervised diarization on the interviews.

Our diarization system is based on i-vectors and agglomerative hierarchical clustering (AHC) and utilizes a diarization system for AMI implemented by Vimal Manohar[1]. Based on the AMI data we train a universal background model (UBM) [22] with 2048 Gaussians, an i-vector extractor [23, 24] with 128 components, and a probabilistic linear discriminant analysis (PLDA) [25] model as an i-vector distance measure.

The diarization process starts by partitioning the recordings into uniform segments of 1.5 seconds with an overlap of 0.75 seconds [26]. Using the VAD system it excludes non-speech frames and extracts one i-vector per segment. The diarization system agglomeratively clusters the i-vectors based on average PLDA scores [26] by merging the clusters with the highest score. Since each ILSE interview has only two speakers, the clustering merges the segments into two clusters.

In addition to speech from both interviewer and participant the recordings contain a variety of artifacts that occur in spontaneous conversations such as crosstalk, back-channeling and

[1] Available online: `https://github.com/vimalmanohar/kaldi/tree/kaldi-diarization-v2/egs/ami/s5b`

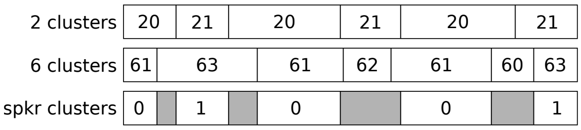

Figure 2: *Cluster Assignment: Each block represents an audio segment assigned to one cluster. Cluster 20 shares most audio time with cluster 61, their overlap is thus assigned to the first speaker cluster (0). Correspondingly, the overlap of clusters 21 and 63 is assigned to the second speaker cluster (1). Gray blocks become discarded segments.*

background noises like paper shuffling. Since we aim for pure speaker segments without any of these artifacts we cluster the i-vectors a second time, this time stopping at six clusters. As a result, artifacts are agglomerated in separate clusters.

3.3. Cluster Assignment

We combine the results of the two 2-cluster and 6-cluster results to obtain pure segments consisting of only participant and interviewer speech (Figure 2). The procedure involves two steps: first, we identify which two clusters most likely contain speech; second we assign one of these two speech clusters to the participant and the other to the interviewer.

The first step is performed by calculating the overlap between audio segments of the 2-cluster diarization and the 6-cluster diarization (see Figure 2). The two cluster pairs that display the largest overlap (i.e. share the most audio duration) are assigned as corresponding speaker clusters. The overlapping segments belonging to these speaker clusters are considered as trustworthy speech segments, all other segments are discarded.

In the second step the larger speech cluster is assigned to the participant, the other cluster is assigned to the interviewer. This simple heuristic is suitable since the ILSE interviews are designed to keep the interviewers' contribution at a minimum. Finally, the segments of the participant cluster are kept for further processing, all other segments are discarded.

We evaluated the speaker diarization and cluster assignment on the same held-out set that was used for the VAD evaluation. In preparation for dementia detection we optimized for speaker error instead of diarization error because incorrectly assigned speech has a much higher impact on the classification result than missed speech. The 2- and 6-speaker diarizations reached a speaker error rate of 18% and 12%, respectively. By considering only trustworthy speaker segments the performance improves to 6% with a loss of about 20% of the speech data.

4. Detecting Dementia

4.1. Acoustic Features for Dementia Detection

Acoustic features capture how speakers talk, instead of what they say. From the realm of acoustic features we select pause-based features: mean, median and variance of the duration of speech and pause segments, percentage of pause time, and pause counts (for a detailed description see Weiner et al. [4]). We use participants' speech segments as well as the non-speech segments between the participant's speech segments as identified in the data preparation stages (Section 3) to capture the occurrence and duration of pauses. Unlike in our previous work [4, 5] we do not use any features that rely on manual or automatic transcriptions [5]. Once automatic transcription is completed, we expect that with a combination of acoustic and linguistic features we can further improve the detection.

4.2. Dementia Detection from Full Interviews

We detect cognitive diagnoses (control vs AACD vs AD) from the interview recordings. For the transcribed data set, we infer participant speech segments from the transcriptions. For the untranscribed data set we run the process described in Section 3.

For each interview we extract our nine features (Section 4.1) from the participant speech, representing each interview by one feature vector, and select features based on mutual information. Finally we train a Gaussian classifier to discriminate the three cognitive diagnoses. The experiments are based on scikit-learn [27].

We train and evaluate the classifier in a leave-one-person-out cross-validation. Each participant contributed to the study 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 classes in ILSE is determined by their natural occurrence [28], UAR is more suitable than a weighted metric such as accuracy. The chance level for a three-class classification is at UAR = ⅓.

On the transcribed data set our experiment achieves a result of 0.493 UAR. The left confusion matrix in Figure 3 shows a reasonably good classification of AACD and AD. Unfortunately, the classifier has a strong bias towards the AACD and AD classes which leads to a poor detection of the control group.

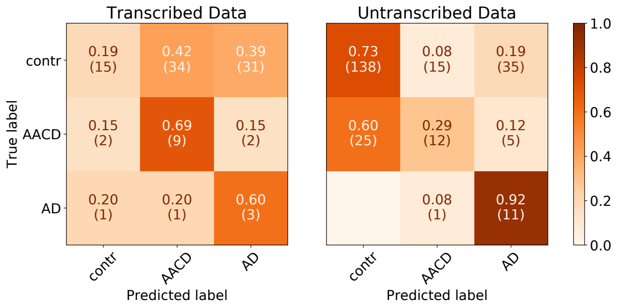

Figure 3: *Confusion matrix of the result on the transcribed (left) and untranscribed (right) data. The number of samples is given in parenthesis.*

For the untranscribed data set the experiment yields UAR = 0.645. This result by far outperforms the result on the transcribed data. The confusion matrices in Figure 3 clearly show this enhanced performance. No participant with AD was incorrectly classified as cognitively healthy which is an improvement over the result on the transcribed data. Furthermore, a good discrimination between healthy people and people affected by AD is exactly what area-wide dementia pre-screening needs. These improvements are most likely due to the larger amount of training data. They also indicate that the unsupervised data preparation process (Section 3) is reliable.

4.3. The Effect of Audio Duration on Dementia Detection

We have shown that we can reliably detect dementia based on automatic unsupervised speaker segmentation, and gain performance improvements by leveraging data from more participants. In a further step we investigate how the total duration of the participants' speech affects dementia detection. This enables us to see how much speech needs to be collected from patients by a dementia pre-screening system.

In the untranscribed data set we then split the participant's speech into non-overlapping segments of equal duration. In this way we extract participant speech segments with a duration of 2.5, 5, 7.5, 10, 12.5, 15, 17.5, and 20 minutes. If the last segment of an interview is shorter than 80% of the target duration, we omit it. In Section 4.2 each feature vector represents per participant the speech of the complete interview. In contrast, in the following experiments, each feature vector represents one segment of speech. For example, in the case of 10-minute speech segments, we have one feature vector for the first ten minutes of participant speech in an interview, one for the second ten minutes of speech in that interview and so forth.

As a result, we have more segments and thus data points for shorter than for longer durations. In order to evaluate the effect of the audio duration on dementia detection without any influence from the amount of training data we employ Monte Carlo sampling with replacement: For each interview we pick the feature vector of one segment uniformly at random so that we always have the same number of training samples: one sample per interview. With these randomly selected feature vectors we perform a leave-one-person-out cross validation to select the best features and train a Gaussian model. We run 1,000 iterations of picking one feature vector per interview and detecting dementia. Finally, we calculate the average and standard deviation UAR (see blue line and error bars in Figure 4). These results show that we can detect dementia from short segments of participant speech. Even for segments as short as 2.5 minutes we achieve an average result of UAR $= 0.552$. The best result is achieved for 12.5-minute segments: the average UAR is 0.597 with a standard deviation of 0.03.

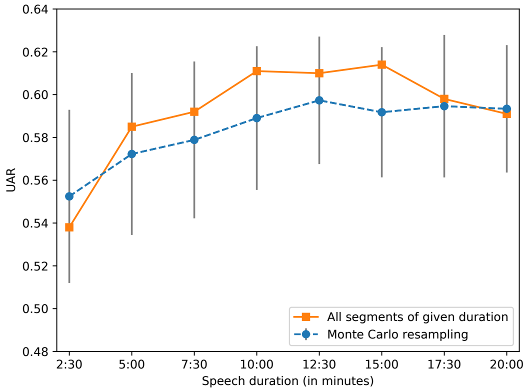

Figure 4: *Average UAR and standard deviation of the 1,000 iterations of Monte Carlo sampling of feature vectors representing different durations of speech (━■━). Results using all segments as training data for different durations of speech (-◆-).*

Going further we leverage the fact that for shorter segment durations more training data are available. For this purpose we now train and evaluate models using all segments for each duration. The orange line in Figure 4 shows these results. For almost all considered durations the UAR results improve when we use all available segments. The trend in the results is clearer in the results using all data than it was in the result of the Monte Carlo resampling: On the one hand the UAR improves as the segment duration increases. Increased segment duration means that each feature vector represents more speech, i.e. more information about the subject is available to the classifier. On the other hand the UAR decreases as segment duration increases further and less training data samples are available. The best trade-off in this dataset between the the amount of information that contributed to a feature vector and the number of training data samples occurs in the range of 10 to 15 minutes.

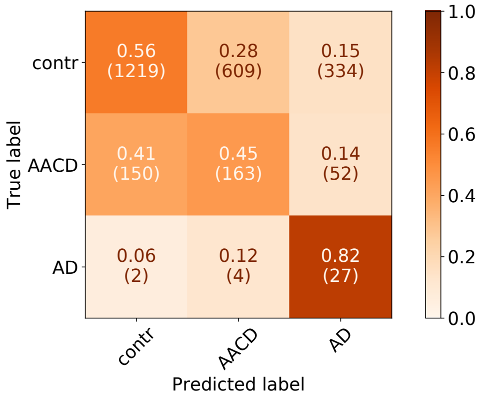

Figure 5: *Confusion matrix of the result when using all the 12.5-minute segments of the untranscribed data set (The number of samples is given in parenthesis).*

For the segment duration of 12.5 minutes we have 365 segments of AACD, 33 AD segments and 2162 segments contributed by control subjects. We use all this available data to train and evaluate models for 12.5-minute segments. The overall result (UAR $= 0.610$) is comparable to the results on the whole interviews (Section 4.2) while using much less audio data per sample. The confusion matrix in Figure 5 clearly shows a high confusability between control and AACD with no bias towards one of the two classes, and again a very good result for AD.

5. Conclusions

We have investigated a very time- and cost-effective approach to fully automatic dementia detection using speech. Using unsupervised speaker diarization we identified participant speech segments in a large data set of 241 interviews from 218 participants. Using acoustic features extracted from these segments we have detected dementia with an UAR of 0.645. Leveraging the large number of recordings this system outperformed a system trained on a smaller data set with supervised speaker segmentation by a large margin. We have further shown that we can detect dementia using speech segments as short as 2.5 minutes, but achieve the best results using segments in the range between 10 and 15 minutes.

6. 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.

S. Umesh would like to thank the Alexander von Humboldt Foundation for supporting his stay with the Cognitive Systems Lab in Bremen.

7. References

[1] 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.

[2] 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.

[3] C. Sattler, H.-W. Wahl, J. Schröder, A. Kruse, P. Schönknecht, U. Kunzmann, T. Braun, C. Degen, I. Nitschke, W. Rahmlow, P. Rammelsberg, J. S. Siebert, B. Tauber, B. Wendelstein, and A. Zenthöfer, *Interdisciplinary Longitudinal Study on Adult Development and Aging (ILSE).* Singapore: Springer, 2017, pp. 1213–1222.

[4] J. Weiner, C. Herff, and T. Schultz, "Speech-Based Detection of Alzheimer's Disease in Conversational German," in *INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association*, 2016.

[5] J. Weiner, M. Engelbart, and T. Schultz, "Manual and Automatic Transcription in Dementia Detection from Speech," in *INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association*, 2017.

[6] 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.

[7] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert, and R. David, "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 1, no. 1, pp. 112 – 124, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352872915000160

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] L. Hernández-Domínguez, E. García-Cano, S. Ratté, and G. Sierra-Martínez, "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*, 2016.

[14] L. Zhou, K. C. Fraser, and F. Rudzicz, "Speech recognition in alzheimers disease and in its assessment," in *INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association*, 2016, pp. 1948–1952.

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

[16] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar, "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," in *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*, 2014, pp. 27–36.

[17] 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.

[18] P. Martin and M. Martin, "Design und Methodik der Interdisziplinären Längsschnittstudie des Erwachsenenalters," in *Aspekte der Entwicklung im mittleren und höheren Lebensalter: Ergebnisse der Interdisziplinären Längsschnittstudie des Erwachsenenalters (ILSE)*, P. Martin, K. U. Ettrich, U. Lehr, D. Roether, M. Martin, and A. Fischer-Cyrulies, Eds. Steinkopff, 2000, pp. 17–27.

[19] 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.

[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*, 2014, pp. 2650–2654.

[21] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The ami meeting corpus: A pre-announcement," in *International Workshop on Machine Learning for Multimodal Interaction.* Springer, Berlin, Heidelberg, 2005, pp. 28–39.

[22] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[23] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[24] S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, 2013.

[25] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on.* IEEE, 2007, pp. 1–8.

[26] G. Sell and D. Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in *Spoken Language Technology Workshop (SLT), 2014 IEEE.* IEEE, 2014, pp. 413–417.

[27] 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.

[28] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language," in *INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association*, 2016.