EXPLORING DEMENTIA DETECTION FROM SPEECH: CROSS CORPUS ANALYSIS

Ayimnisagul Ablimit[1]*, *Catarina Botelho*[2]*, *Alberto Abad*[2], *Tanja Schultz*[1], *Isabel Trancoso*[2]

[1]Cognitive Systems Lab, University of Bremen, Germany
[2]INESC-ID/Instituto Superior Técnico, University of Lisbon, Portugal

ABSTRACT

In this work, we present a qualitative and quantitative analysis of speech and language features derived from two different corpora with the aim to predict early signs of dementia. One corpus consists of the Interdisciplinary Longitudinal Study on Adult Development and Aging (ILSE) designed to investigate satisfying and healthy aging. It consists of more than 6500 hours of biographic interviews from 1000 participants recorded over the course of 20 years. The other corpus is a cross sectional data set created for the ADReSS challenge 2020. In an experimental study we describe a large variety of acoustic and linguistic features that are automatically extracted from speech and corresponding transcriptions. We compare different traditional classifiers, i.e. Gaussian Mixture Models, Linear Discriminant Analysis, and Support Vector Machines. Our final performance results surpass the ADReSS benchmarks.

Index Terms— Speech & language, Alzheimer's disease, acoustic and linguistic features, ILSE corpus, ADReSS challenge

1. INTRODUCTION

Fifty-five million people worldwide live with dementia – a syndrome that leads to the deterioration of the cognitive function beyond what is considered normal in biological aging [1]. The most frequent form of dementia, Alzheimer's disease (AD), is a progressive neurodegenerative disorder which causes impairments in memory, spatio-temporal orientation, and language. Population aging is responsible for an increase of new AD cases, and creates the need for scalable, cost-effective methods that are able to detect early stage AD. Speech and language biomarkers are strong indicators of dementia, and provide a low-cost and widespread alternative for the assessment of cognitive states. Several researchers have proposed methods to detect dementia and AD from speech and language features (e.g. [2, 3, 4, 5]), making use of different speech corpora and achieving promising results. Nevertheless, corpora are often very limited in terms of size and population representation, which poses questions on the generalizability of the results. In fact, very few studies perform cross-corpora comparisons. Arguably, the uncertainty about the generalizability of results and methods is one of critical issues to solve before deploying speech and machine learning based solutions in clinical applications.

Our recent work [6, 7, 8] has focused on the detection of dementia from speech recorded over biographic interviews acquired within the *Interdisciplinary Longitudinal Study on Adult Development and Aging* (ILSE) – a study that aims to investigate satisfying and healthy aging in middle adulthood and later life [9]. The ILSE corpus is fairly large, when compared to other health-related speech corpus, as it consists of approximately 6,500 hours of recordings, including over 1000 participants, recorded over a 20 year period. Despite the limited number of participants with AD, which is a result of the natural prevalence of dementia in the population studied, the corpus has a very large potential, as it enables aging-related research in many disciplines including geriatrics, psychology, ecological, gerontology, sociology, history, and linguistics [10]. One of the main challenges of this corpus is the changing recording conditions over time.

Due to the coverage of sensitive private information, the corpus is not publicly available. Thus, in this work we propose to analyse dementia detection in parallel, for two corpora: the imbalanced and longitudinal conversational ILSE corpus, and the balanced subset of Dementia Bank [11] used in the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) challenge [12], which is a publicly available corpus of spontaneous speech samples from a picture description task. We explore the audio and text modalities using 8 distinct sets of features/embeddings. We further discuss the longitudinal dimension of ILSE corpus. This initial study, focusing on a subset of ILSE corpus, will set the ground and necessary baselines before diving into larger experiments involving the entire ILSE corpus. In summary, this cross-corpus and cross-lingual study aims at answering the following key questions:

1. Are the distributions of features considered informative for the detection of AD similar in the two distinct corpora?
2. Do the best performing methods for dementia detection in the ILSE corpus work in the Dementia Bank?
3. What aspects should be taken into special consideration when performing this analysis in a longitudinal corpus versus a cross-sectional corpus?

2. RELATED WORK

The ADReSS 2020 challenge [12] released a benchmark speech dataset balanced in terms of age and gender, for two tasks: AD speech classification, and neuropsychological score regression. The main goal of this challenge was to address the lack of standardisation that currently affects the field, and introduce a dataset on which the different approaches could be systematically compared. The ADReSS contains two baselines for AD classification. The acoustic baseline uses ComParE features [13] and LDA classifier, achieving 62.5% accuracy; the linguistic baseline uses a set of 34 language outcome measures (e.g., duration, mean length of utterances, type-token ratio, open-closed class word ratio, percentages of 9 parts of speech) and LDA classifier, and achieved an accuracy of 75%. The winners of the ADReSS 2020 challenge achieved an accuracy of 89.6% [14] and 85.45% [15], using acoustic and text-based features.

*Both authors contributed equally to this work. [1]This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through the project "ALMED - acoustic and linguistic features for early prediction of cognitive deficits" (403605461). [2]This work was supported by national funds through *Fundação para a Ciência e a Tecnologia* (FCT) under projects UIDB/50021/2020, CMU/TIC/0069/2019, and grant number SFRH/BD/149126/2019.

3. CORPORA

3.1. ILSE

The ILSE longitudinal corpus includes participants born in Germany, in two distinct cohorts: 1930-1932 and 1950-1952. Each participant engaged in up to four measurements, conducted in intervals of approximately 5 years. At each measurement, among other tasks, the subjects were recorded during biographic interviews, where they gave elaborate answers to open-ended questions. The average duration of the recordings became shorter with each follow-up interview, given the accumulated data. The speech recordings of the first two measurements were stored on tapes. From the third measurement, digital recording devices were used. All interviews have been digitized using a sampling rate of 16kHz, and 16-bit linear PCM quantization. Further details about the ILSE corpus can be found in [10].

The effort to manually transcribe the interviews is ongoing. Currently, we have manual transcripts for 145 interviews from 91 participants. These transcripts have varying quality and do not include time alignment information. To distinguish this subset of the ILSE corpus from the subsets used in previous and future works, we will denote this subset as $\text{ILSE}_{\text{m145}}$. Table 1 shows the number of interviews and duration of the dataset. Although $\text{ILSE}_{\text{m145}}$ includes interviews associated with three diagnoses: controls, AD and age-associated cognitive decline (AACD), we only conduct the classification experiments on AD versus controls, to allow a fair comparison with the ADReSS corpus (see section 3.2).

Table 1. $\text{ILSE}_{\text{m145}}$: number and total duration [hours] of the interviews per diagnosis class.

	control	AD	AACD
Interviews	108	16	21
Duration of Interviews	405.51	21.29	57.82
Participant's speech	270.27	12.95	39.53

3.2. ADReSS

The ADReSS corpus [12] contains speech recordings and the annotated transcriptions of 156 subjects (78 have AD, and 78 are healthy controls matched for age and gender) describing the Cookie Theft picture from the Boston Diagnostic Aphasia Examination [16].

Speech recordings were segmented using Voice Activity Detection (VAD) and later normalised. The dataset made available contained both full enhanced audio, and normalised audio chunks. In our approach, we used the full enhanced audio and the corresponding transcriptions. We divided data into training / test sets with 108/48 subjects, respecting the partitions proposed in [12].

4. METHOD

4.1. Pre-processing

ILSE data Since the interviews have not been segmented at speaker turns, speaker diarization [17] is applied to exclude speech from interviewers. The manual transcripts include speaker-turn annotations, although no time alignment information is provided.

ADReSS data For each recording, we discarded the interventions of the interviewer, and processed only the participants' parts. The transcriptions were simplified by removing annotations of complex events, and considering only plain text. This approximates the ADReSS transcriptions to the transcriptions available for ILSE, and to the output that can be generated by an ASR system.

4.2. Automatic Extraction of Features for Screening

We extracted speech and linguistic indicators to capture speaker characteristics as well as the content and form of the message being transmitted. The 8 sets of features described below were extracted at the interview level.

Linguistic Inquire and Word Count (LIWC) was developed as a tool for psychologists, and has been applied to mark individual differences in cognitive processing. In our prior work [18], we have used LIWC for dementia screening and achieved promising results. We map each word to one of N categories ($N = 68$ for German and $N = 64$ for English), using a language-dependent LIWC-dictionary, and then we compute the percentage of occurrences of each category in the transcripts. The percentage of occurrences of each category is one feature. LIWC-categories may differ slightly across languages, but for each language they were created to categorize words, capture various social, cognitive and affective processes.

Part-of-Speech (PoS) PoS tags groups words with similar grammatical properties into the same PoS tag, capturing the grammatical structure of participants' speech. The transcripts are tagged using a language-dependent TreeTagger [19, 20], and each feature corresponds to the percentage of occurrences of each tag. The dimension of this feature set is 55 for ADReSS, 57 for ILSE.

Perplexity The perplexity reflects how easy or hard it is to predict the words in a text. Given that subjects with cognitive deficits use smaller vocabulary and simple sentences when communicating, it is expected that their speech is easier to predict, and thus the perplexity is lower. For each interview, we developed n-gram language models by taking 80% of the segments of the interview as training set and evaluating the perplexity of the developed language models on the remaining 20% (n = 1, 2 and 3). The set of features related to the language models has dimension 15 [21].

Part-of-Speech (PoS) Perplexity We also apply perplexity to the PoS tags to explore the complexity of grammatical usage, using a 5-gram language model. This feature set has dimension 23 [21].

Voice activity detection based (VAD) features are a set of 11 clinically interpretable features based on silence/pauses metrics used in [17]. This set includes silence count, silence count ratio, mean silence duration, variance silence duration, median silence duration, mean speech duration, variance speech duration, median speech duration, silence to speech ratio, mean silence count, and silence rate.

ComParE The ComParE 2013 [13] feature set consists of 6373 features, including energy, spectral, MFCC, and voicing related low-level descriptors, and statistical functionals.

i-vectors [22] model total speaker and channel variability, and have been shown to also carry information about the health status of the speaker [23]. The 128 dimensional i-vectors were extracted with Kaldi [24], using the i-vector extractor trained on the AMI data set.

ECAPA-TDNN embeddings [25] are deep neural network based embeddings that achieve the state-of-the-art in speaker recognition, and correspond to an enhanced version of the x-vector architecture [26]. X-vectors have been shown to carry paralinguistic information. In particular, they have been used for the automatic detection of different diseases, including AD [27, 28]. ECAPA-TDNN embeddings, with 192 dimensions, were extracted using the model made available by SpeechBrain [29], pre-trained on Voxceleb data[1].

[1] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

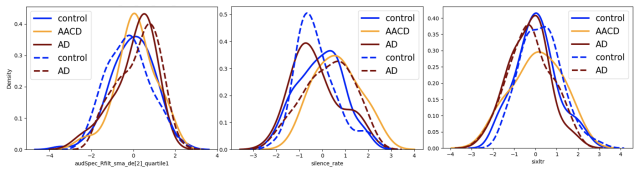

Fig. 1. Density plots for feature distribution across both datasets. Left: *audSpec_Rfilt_sma_de[2]_quartile1* from ComParE feature set; center: silence rate from VAD feature set; right: "six+ letter words" from LIWC feature set. The solid lines refer to ILSE and dashed lines refer to ADReSS.

The models used for extracting i-vectors and ECAPA-TDNN embeddings were not re-trained specially for ILSE using German data, since (1) to the best of our knowledge, there is no comparable German dataset regarding data size and number of speaker, and (2) Studies have successfully used x-vectors trained with VoxCeleb for disease detection in a different language [30, 31]

4.3. Feature normalization

All feature sets except i-vectors and ECAPA-TDNN embeddings were normalized with zero-mean and unit-variance normalization. For $\text{ILSE}_{\text{m145}}$, the normalization was performed separately for the interviews that belong to the same measurement time. The rationale behind this choice is that the interviews at different measurement times are very distinct in terms of recording conditions, length (varies from roughly 1h to 5h per interview, excluding the interviewer segments) and topics discussed. Thus, this type of normalization was chosen to avoid the possible bias introduced by the fact that the majority of control interviews are from the first measurement period, and all the AD interviews are from the third one. i-vectors were scaled such that the L2-norm of each vector was 1.

4.4. Classification

The AD screening consists of creating a binary classifier that distinguishes AD and controls. The classification is performed on each feature set separately. We compare three different classifiers: Gaussian Mixture Model (GMM) with diagonal covariance, Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM) with Radial Basis Function (RBF). We train separate models for each dataset, with leave-one-subject-out cross validation, to cope with the limited size of the datasets. For the ADReSS data, we also evaluate on the held-out test set. The performance is evaluated in terms of unweighted average recall (UAR). For ADReSS, because it is a balanced dataset, UAR is equivalent to accuracy, thus allowing a direct comparison with other works that report accuracy.

5. RESULTS

5.1. Feature distribution

We analyzed qualitatively the distributions of the features for both datasets, using density plots. Figure 1 shows only some examples of the feature distributions, for the sake of conciseness and space. For the *audSpec_Rfilt_sma_de[2]_quartile1* from the ComParE feature, we observe that the control distribution and the AD distribution are slightly shifted from each other, in the same direction for both datasets – which is a desired behaviour for features that could generalize AD detection across different corpora. On the other hand, the same does not happen for the features "sixltr" (six or greater letter words) category in LIWC and silence rate from VAD feature set.

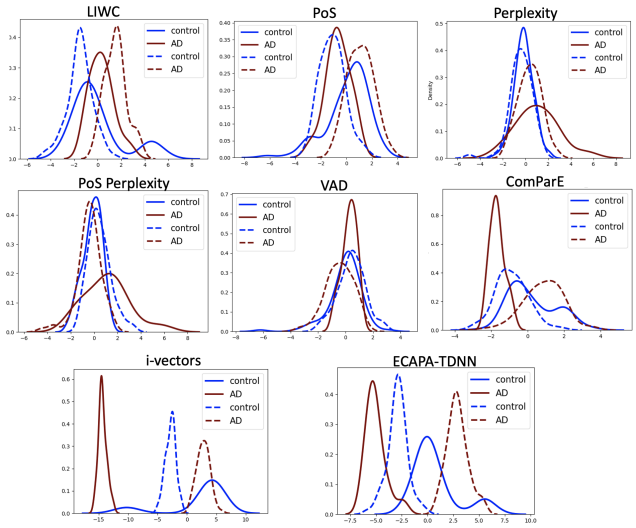

Fig. 2. LDA projections of each feature set. The solid lines refer to ILSE and dashed lines refer to ADReSS.

For the silence rate, the distribution shifts seem to be inverted: ADs seem to have a higher silence rate when compared to controls in the ADReSS corpus, and a lower silence rate in ILSE corpus. Regarding the "sixltr", we observe that the AD and control distributions are shifted from each other in the case of ADReSS data, but not in the case of ILSE data. This is probably due to the fact that 6-letter words are considered large in English, and therefore more typical of formal speech [32], thus it is understandable that people with cognitive deficits would use less of these words, when compared to healthy people. The same does not happen in German, which typically uses larger words - thus "sixltr" may not be a strong indicator of cognitive impairment. However, the feature "sixltr" has been used in previous studies in German [33, 18]. These results call for future research to identify which features can be transferred across different languages and speech tasks, and how to adapt the features to make them meaningful across domains.

The large number of individual features does not allow an individual analysis of all of them, thus figure 2 shows the one-dimensional LDA projection of each feature set. LDA projections are supervised (*control* vs. *AD*), and they were computed independently for each dataset, using all data available. Through visual inspection, one can observe that the projections of ECAPA-TDNN embeddings and i-vectors seem to have a smaller overlap between control and AD distributions, than the rest of the feature sets.

Besides the comparison of features across corpora, we also explored qualitatively how speaker embeddings vary across different interviews in $\text{ILSE}_{\text{m145}}$. For this end, we plotted in Figure 3 t-SNE projections [34] of ECAPA-TDNN embeddings extracted at the interview level for all the 37 subjects that have two or more interviews. Through visual inspection, we observe that the embeddings extracted from different interviews of the same subject sometimes (roughly for half of the subjects) lie in the same region of space, and sometimes are plotted far apart. This observation is not stronger in the cases where subjects develop AD or AACD, as it is possible to make the same observation (with roughly the same frequency) in healthy subjects. We thus hypothesize that the variability for the same speaker embeddings across different interviews is probably explained by larger margin through changes in the recording conditions and/or aging, which may obfuscate more subtle changes due to AD.

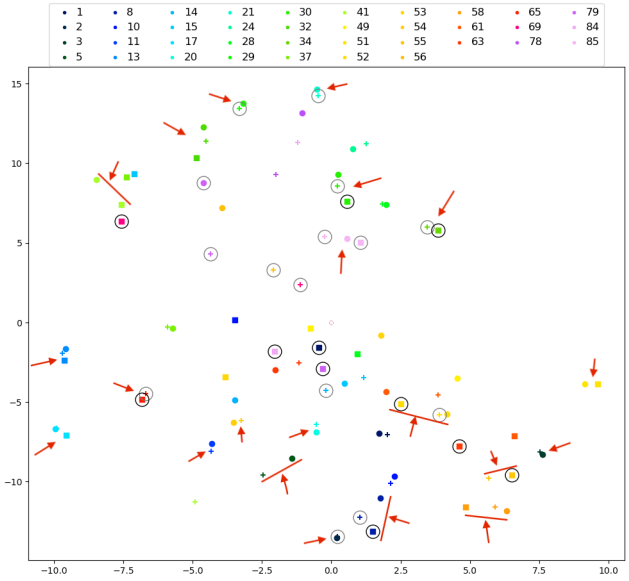

Fig. 3. t-SNE representations of ECAPA-TDNN embeddings extracted at the interview level for all the 37 subjects in $ILSE_{m145}$ that have 2 or more interviews. Each subject is represented by a different color, and •, +, and ■ denote interviews that occurred at the first, second and third round of interviews, respectively. Black circles around the marks denote an interview where the subject has AD and gray circles denote an interview where the subject has AACD. Red arrows highlight subjects for which all their interviews were projected to the same region – showing a purely qualitative analysis.

5.2. Classification

Tables 2 and 3 summarize the results obtained when performing the classification using each of the feature sets, for ADReSS and ILSE, respectively. For ADReSS, the acoustic features were able to achieve 66.7% of UAR, and the linguistic features achieve 77.1% of UAR on the held-out test set. These results surpass the baseline provided for the 2020 ADReSS challenge [12] (62.5% and 75%, respectively). The linguistic features which are more informative are LIWC and PoS tags, while ECAPA-TDNN embeddings achieve the best performance when compared to other acoustic representations. SVM is the best classifier. Furthermore, we observe a large performance gap between training cross validation and the held-out test set, which reflects some overfitting during training. Nevertheless, the models are still able to surpass baseline performance on the test set.

For the ILSE corpus, because we do not have defined a held-out test set, we report all results only in terms of leave-one-subject-out cross validation. The best results for the linguistic features (83.8% UAR) are achieved with the pair PoS tag features and GMM classifier, while the best result achieved with acoustic features is 86% with i-vectors and SVM classifier.

In general, when comparing the cross-validation results of ILSE with the cross-validation results of ADReSS, the former appear worst than the latter. Nevertheless, it is possible that the models are overfitting more the ADReSS data. Another important aspect to highlight is the fact that the best classifiers and the best feature sets for one dataset do not match the best for the other dataset, with the exception of the PoS tag features.

We also experimented with models trained in one dataset and evaluated on the other, i.e., we took a model trained on ILSE, and tested on ADReSS, and vice-versa. The results were barely above chance level. The exception occurred for the perplexity features, that achieved a UAR of 62.5% on the test set of ADReSS, using an SVM model trained in ILSE.

Table 2. Classification results in ADReSS (UAR).

	GMM		LDA		SVM	
	CV	test	CV	test	CV	test
LIWC	69.4	60.4	99.1	68.8	99.1	**77.1**
PoS	64.8	54.2	89.8	68.8	92.6	**77.1**
Perplexity	55.6	54.2	68.5	45.8	73.1	56.2
PoS Perplexity	55.6	47.9	71.3	58.3	68.5	54.2
VAD	50.9	56.2	69.4	60.4	70.4	62.5
ComParE	94.4	47.9	75.0	60.4	99.1	58.3
i-vectors	55.6	50.0	60.2	65.5	56.5	54.2
ECAPA-TDNN	93.5	**66.7**	95.4	60.4	99.1	**66.7**

Table 3. Classification results in ILSE (UAR).

	GMM	LDA	SVM
LIWC	78.9	39.2	48.1
PoS	**83.8**	41.2	38.0
Perplexity	55.9	55.8	71.6
PoS Perplexity	55.7	54.9	50.0
VAD	45.7	49.5	56.7
ComParE	49.5	69.0	34.0
i-vectors	84.6	79.3	**86.0**
ECAPA-TDNN	61.9	66.7	52.9

One should consider that the two datasets differ not only in language, but also in the number of subjects, total duration of the audio, and type of data. ILSE recordings correspond to German biographic interviews, while the ADReSS recordings correspond to the English description of the Cookie Theft. While it is true that both tasks trigger spontaneous speech, the area and diversity of vocabulary, as well as the emotional content are expected to differ across datasets. These facts may explain the different results found across datasets. Nevertheless, considering that datasets for medical diagnosis are typically small and heterogeneous, we argue that it is crucial that future research discusses how to translate results across different domains, and how to measure robustness and trust of the results achieved.

6. CONCLUSION

This work explores the screening of dementia from speech in a longitudinal conversational corpus – the ILSE corpus – and establishes a comparison with a publicly available cross-sectional corpus – the ADReSS corpus. This is an initial study, designed to establish baselines and set the grounds for future experiments. We discuss the feature distribution for both corpora, including the aspects that could be a source of bias in a longitudinal corpus, such as ILSE. The classification experiments achieve promising results, and surpass the baselines proposed for the ADReSS challenge. Nevertheless, the fact that the best classifiers and the best feature sets for one dataset do not match the best for the other dataset, raises questions on the transferability of the methods to new domains, such as different languages, different tasks and different recording conditions. Hence, it calls for future research to explore, on one hand, how can we be sure that what we are learning is indeed attributable to the disease and not to aging, differences in recording conditions or other characteristics of the dataset; and on the other hand, which methods can we use to leverage models trained in one domain to perform in a new domain.

References

[1] W. H. Organization, "Dementia," https://www.who.int/news-room/fact-sheets/detail/dementia, September 2021 (accessed on September 28, 2021).

[2] W. Jarrold et al., "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," in *CLPsych*, 2014.

[3] L. Tóth, G. Gosztolya, V. Vincze, I. Hoffmann, and G. Szatlóczki, "Automatic detection of mild cognitive impairment from spontaneous speech using asr," in *INTERSPEECH*. ISCA, 2015.

[4] J. Weiner, C. Herff, and T. Schultz, "Speech-Based Detection of Alzheimer's Disease in Conversational German," in *INTERSPEECH*, 2016.

[5] A. Pompili, A. Abad, D. M. de Matos, and I. P. Martins, "Pragmatic aspects of discourse production for the automatic identification of alzheimer's disease," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 261–271, 2020.

[6] J. Weiner, C. Frankenberg, J. Schröder, and T. Schultz, "Speech reveals future risk of developing dementia: Predictive dementia screening from biographic interviews," in *ASRU*, 2019.

[7] A. Abulimiti, J. Weiner, and T. Schultz, "Automatic speech recognition for ilse-interviews: Longitudinal conversational speech recordings covering aging and cognitive decline.," in *INTERSPEECH*, 2020.

[8] A. Ablimit and T. Schultz, "Automatic Speech Recognition for Dementia Screening using ILSE-Interviews," in *ITG Speech Communication*, 2021.

[9] C. Sattler, H.-W. Wahl, J. Schröder, A. Kruse, P. Schönknecht, U. Kunzmann, and A. Zenthöfer, "Interdisciplinary longitudinal study on adult development and aging (ILSE)," *Encyclopedia of geropsychology*, pp. 1–10, 2015.

[10] P. Martin, M. Grünendahl, and M. Schmitt, "Persönlichkeit, kognitive leistungsfähigkeit und gesundheit in ost und west: Ergebnisse der interdisziplinären längsschnittstudie des erwachsenenalters (ilse)," *Zeitschrift für Gerontologie und Geriatrie*, vol. 33, no. 2, pp. 111–123, 2000.

[11] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[12] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: the adress challenge," *INTERSPEECH*, 2020.

[13] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *MM*, 2013.

[14] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease.," in *INTERSPEECH*, 2020.

[15] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, "Automated screening for Alzheimer's dementia through spontaneous speech.," in *INTERSPEECH*, 2020.

[16] H. Goodglass, E. Kaplan, and S. Weintraub, *BDAE: The Boston Diagnostic Aphasia Examination*, Lippincott Williams & Wilkins Philadelphia, PA, 2001.

[17] J. Weiner, C. Herff, and T. Schultz, "Speech-based detection of Alzheimer's disease in conversational german.," in *INTERSPEECH*, 2016.

[18] J. Weiner and T. Schultz, "Automatic Screening for Transition into Dementia using Speech," in *ITG*, 2018.

[19] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *New methods in language processing*, 2013, p. 154.

[20] H. Schmid, "Improvements in part-of-speech tagging with an application to german," in *Natural language processing using very large corpora*, pp. 13–25. Springer, 1999.

[21] C. Frankenberg et al., "Perplexity – a new predictor of cognitive changes in spoken language? – results of the Interdisciplinary Longitudinal Study on Adult Development and Aging (ILSE)," *Linguistics Vanguard*, vol. 5, 06 2019, s2.

[22] 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, 2010.

[23] Y. Hauptman et al., "Identifying distinctive acoustic and spectral features in parkinson's disease.," in *INTERSPEECH*, 2019.

[24] D. Povey et al., "The kaldi speech recognition toolkit," in *ASRU*, 2011.

[25] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *INTERSPEECH*, 2020.

[26] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*, 2018.

[27] S. Zargarbashi and B. Babaali, "A multi-modal feature embedding approach to diagnose alzheimer disease from spoken language," *arXiv preprint arXiv:1910.00330*, 2019.

[28] A. Pompili, T. Rolland, and A. Abad, "The INESC-ID multimodal system for the ADReSS 2020 challenge," in *INTERSPEECH*, 2020.

[29] M. Ravanelli et al., "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[30] L. Moro-Velazquez, J. Villalba, and N. Dehak, "Using x-vectors to automatically detect parkinson's disease from speech," in *ICASSP*, 2020.

[31] L. Jeancolas et al., "X-vectors: New quantitative biomarkers for early parkinson's disease detection from speech," *Frontiers in Neuroinformatics*, vol. 15, pp. 4, 2021.

[32] V. Marian, J. Bartolotti, S. Chabal, and A. Shook, "Clearpond: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities," *PLoS ONE*, 2012.

[33] M. Wolf, A. B. Horn, M. R. Mehl, S. Haug, J. W. Pennebaker, and H. Kordy, "Computergestützte quantitative textanalyse: äquivalenz und robustheit der deutschen version des linguistic inquiry and word count," *Diagnostica*, vol. 54, no. 2, pp. 85–98, 2008.

[34] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.