The Influence of Automatic Speech Recognition on Linguistic Features and Automatic Alzheimer's Disease Detection from Spontaneous Speech

Jonathan Heitz[1,3,4], Gerold Schneider[2], Nicolas Langer[1,3]
[1] Department of Psychology, University of Zurich, Methods of Plasticity Research, Zurich, Switzerland
[2] Department of Computational Linguistics, University of Zurich, Zurich, Switzerland
[3] University Research Priority Program (URPP) Dynamic of Healthy Aging, Zurich, Switzerland
[4] Language & Medicine Competence Centre, University of Zurich, Zurich, Switzerland
jonathan.heitz@uzh.ch, gschneid@cl.uzh.ch, n.langer@psychologie.uzh.ch

Abstract

Alzheimer's disease (AD) represents a major problem for society and a heavy burden for those affected. The study of changes in speech offers a potential means for large-scale AD screening that is non-invasive and inexpensive. Automatic Speech Recognition (ASR) is necessary for a fully automated system. We compare different ASR systems in terms of Word Error Rate (WER) using a publicly available benchmark dataset of speech recordings of AD patients and controls. Furthermore, this study is the first to quantify how popular linguistic features change when replacing manual transcriptions with ASR output. This contributes to the understanding of linguistic features in the context of AD detection. Moreover, we investigate how ASR affects AD classification performance by implementing two popular approaches: A fine-tuned BERT model, and Random Forest on popular linguistic features. Our results show best classification performance when using manual transcripts, but the degradation when using ASR is not dramatic. Performance stays strong, achieving an AUROC of 0.87. Our BERT-based approach is affected more strongly by ASR transcription errors than the simpler and more explainable approach based on linguistic features.

Keywords: Automatic Speech Recognition, Alzheimer's Disease (AD), Automatic AD classification, Stability of Linguistic Features

1. Introduction

Alzheimer's disease (AD) represents a major and rapidly growing burden to the healthcare and economic system (Winblad et al., 2006). The current state of research indicates that therapies need to be administered as early as possible to be effective (Arvanitakis et al., 2019). Therefore, there is an urgent need for accelerating biomarker discovery for AD. However, prevailing biomarkers for AD diagnosis, including genetic testing, CSF, structural MRI and PET imaging, can only be applied to relatively small sample sizes due to their limited availability, excessive costs and invasive nature (Kourtis et al., 2019). This prevents adoption of current biomarker testing in large epidemiological studies, which are imperative to identify the intra-individual progression from healthy to pathological cognitive aging (e.g. AD). Thus, novel non-invasive and inexpensive biomarkers are urgently required to be administered at a large scale with the aim of identifying individuals with indications of AD. Speech and language changes have been identified as early symptoms of AD (Calzà et al., 2021), and their detection and analysis have the potential to be used in large epidemiological studies as a real-time and non-invasive diagnostic method.

There have been various approaches to automatic AD detection from spontaneous speech, most commonly based on audio recordings of a picture description task. Information can be extracted from the raw audio signal (acoustic), the analysis of its transcriptions (linguistic), or a combination of both. Acoustic markers include the presence of pauses, jitter, and shimmer. Linguistic analysis includes the extraction of lexical, syntactic, and semantic features, which have proven to be more informative for AD detection than acoustic features (Cummins et al., 2020). As a result, most research has focused on transcriptions rather than directly on the audio signal. The majority of studies have been based on *manual* transcriptions, requiring manual efforts prohibitive for large-scale adoption. Therefore, automatic speech recognition (ASR) is a prerequisite for an automated real-life AD screening tool.

However, it is unclear how ASR systems behave on data from Alzheimer's patients, as available ASR systems are mostly trained on and optimized for healthy and fluent speech. In addition, it is unclear how transcription errors resulting from the use of ASR affect linguistic features, and how this, in turn, affects downstream AD classification performance.

In this paper, we evaluate three popular ASR systems on the ADReSS dataset (Luz et al., 2020): First, we directly measure the difference between automatically generated transcriptions and their manual counterparts in terms of Word Error Rate

LREC-COLING 2024, pages 15955–15969

(WER), and analyze potential effects of diagnosis, age, and gender on error rates. Second, we assess how stable popular well-established linguistic features are when replacing manual transcripts with ASR output. Last, we evaluate how ASR transcriptions affect the downstream performance of two popular machine-learning approaches, namely i) a fine-tuned BERT model for classification, and ii) a Random Forest model on linguistic features.

2. Related Work

In a general setting, ASR systems have been compared using multiple benchmarks, such as Hugging Face's End-to-end Speech Benchmark (ESB) (Gandhi et al., 2022) or CEASR (Ulasik et al., 2020), aiming to assess ASR's generalizability across different healthy speech datasets. For AD, few studies have compared different ASR systems, including Pan et al. (2021); Li et al. (2022); Syed et al. (2021b); Tang et al. (2023). However, they often use relatively old or custom ASR systems, lagging behind recent developments on the ASR market. For example, no study in this field has evaluated the recently introduced Google "Chirp" model (Zhang et al., 2023), which we use in this study. Importantly, mostly studies did not investigate the effect of age, gender, and diagnosis on error rates of ASR systems.

There have been various approaches of using spontaneous speech for classification of AD. Among the first, Fraser et al. (2016) used manual transcripts to investigate linguistic features and their differentiability for AD, while Luz et al. (2018) presented a system for AD detection based on linguistic features on manual transcriptions. The introduction of the ADReSS dataset (Luz et al., 2020) in 2020 triggered a significant burst of research on the topic. Most research on this dataset found linguistic features more useful than acoustic features (e.g. Cummins et al., 2020), and proposed systems based on manual transcriptions, with influential examples including Balagopalan et al. (2020); Yuan et al. (2020); Syed et al. (2021a); Martinc et al. (2021).

Recently, the use of ASR in AD classification pipelines has become more popular. Some studies use ASR to replace or compensate for missing manual transcriptions (especially studies based on the ADReSSo challenge dataset (Luz et al., 2021), which lacks manual transcriptions). Others compare how AD detection performance differs when using ASR vs. manual transcriptions. In recent AD classification approaches, the frequent use of BERT (Devlin et al., 2018) is apparent, either used as an embedding layer to a downstream classification algorithm (e.g. Syed et al., 2021a; Ilias et al., 2023; Roshanzamir et al., 2021) or directly fine-tuned for classification (e.g. Balagopalan et al., 2020; Yuan et al., 2020; Pan et al., 2021). A more classical yet more interpretable approach to AD classification involves the explicit extraction of features (potentially using prior knowledge on AD symptoms) in combination with general-purpose classification algorithms such as Support Vector Machines (SVM) or Random Forest (RF).

An early comparison of using ASR vs. manual transcripts for AD detection was performed by Weiner et al. (2017), using a limited set of linguistic features and a private German dataset, and finding that some features provide better diagnostic value when calculated on an ASR transcription. Wang et al. (2022) created BERT and RoBERTa embeddings based on transcriptions from a custom ASR on the ADReSS dataset, and used these to train an SVM classifier. Their BERT-based model performs worse using ASR than manual transcripts, while combinations of BERT and RoBERTa profit from using ASR. Importantly, they did not remove interviewer utterances from the audio recordings before running ASR, a shortcoming that is common in prior approaches using the ADReSS dataset. Their models might thus learn from the content or frequency of interviewer interactions, casting doubts on their generalizability. Li et al. (2021) compared different features and classifiers for AD detection, including BERT embeddings. They report similar performance when using ASR than when using manual transcripts. Li et al. (2022) fine-tuned a BERT model on the ADReSS dataset, finding that ASR-generated transcripts allow classification performance similar to manual transcripts, and that improved ASR WER does not produce better classification performance. They did not study the effect on linguistic features, as we do in the following. Soroski et al. (2022) evaluated Google Speech ASR on a private dataset, finding that manual transcripts lead to significantly better results than automatic transcripts. Also, they found that healthy controls exhibit lower error rates than AD patients.

In conclusion, prior research has compared ASR systems for AD spontaneous speech, but mostly on older ASR systems and often failing to analyze potential confounding effects of diagnosis, age, and gender. Numerous studies have proposed AD classification systems based on spontaneous speech. These studies have inspired our work in a) the choice of classification approaches, b) the idea of comparing AD classification performance based on ASR-generated transcriptions against manual transcriptions, and c) the selection of linguistic features. However, to the best of our knowledge, no prior study has evaluated the impact of ASR on the stability of linguistic features

and the consequences of this stability on AD detection. In addition, we have observed the frequent and problematic presence of interviewer utterances in the data.

3. Methodology

3.1. Dataset

We use the ADReSS dataset (Luz et al., 2020) for our experiments. It is balanced for diagnosis, age, and gender (see Table 1) and consists of 156 audio recordings where participants describe the Cookie Theft picture from the Boston Diagnostic Aphasia Examination (Goodglass et al., 2001) in English. Each recording is accompanied by a manual transcription in the CHAT transcription format (MacWhinney, 2000). Metadata for each participant includes age, gender, and Mini Mental State Examination (MMSE) scores.

Some of the audio recordings include interviewer prompts, such as "*Is there anything else?*". These interviewer interactions appear more frequently in AD patients than in control participants. However, they are task-specific, and we want to prevent our models from learning from these interactions. Therefore, we remove all interviewer sections from the audio, leaving only participant speech. We do this using timestamps encoded in the manual transcription files. Similarly, we keep only the participants' utterances from the manual transcripts, and preprocess them by removing special transcription codes such as error or retraction markers. The details of the preprocessing steps are provided in Appendix A.

Table 1 displays basic dataset characteristics for AD and control subjects.

	N (female, male)	Age	Transcript Length	MMSE
AD	78 (43, 35)	66.6 ± 7	99 ± 65	17.8 ± 5.5
Control	78 (43, 35)	66.3 ± 7	116 ± 67	29.0 ± 1.2

Table 1: Characteristics of the ADReSS dataset for AD and control group. Mean and standard deviation are reported for age, MMSE and length of manual transcription (in number of words).

3.2. Automatic Speech Recognition (ASR)

We compare three pre-trained state-of-the-art ASR systems:

1. Wave2vec2 (Baevski et al., 2020), a popular open source system.
2. Whisper (Radford et al., 2023), a recent robust system pre-trained by OpenAI. This is motivated by its popularity.
3. The recently introduced commercial Google Speech "Chirp" model (Zhang et al., 2023), available via a Google Cloud API. This is motivated by the reported excellent performance of the system.

3.3. Classification approaches for Alzheimer recognition

To evaluate the effect on downstream classification performance, we implemented two approaches of binary classification between AD and control. They represent two classes of approaches frequently used in prior research for this task.

Fine-tuned BERT model: We fine-tune a pre-trained $\text{BERT}_{\text{BASE}}$ model[1] (Devlin et al., 2018) with a randomly initialized sequence classification head. We use the default tokenizer, a batch size of 8, a learning rate of $4e-6$, and fine-tune for 30 epochs. These hyperparameters have been determined using hyperparameter testing on the manual transcriptions. Commonly, fine-tuning BERT on a small dataset results in significantly different models when using different random seeds for initialization (Zhang et al., 2020; Dodge et al., 2020). To deal with this, we fine-tune 8 identical models for each setting, differing only in their random seed. For each test sample, one prediction is produced by averaging the individual predictions of these 8 models. This is similar to how prior approaches have dealt with this problem (Yuan et al., 2020; Balagopalan et al., 2020; Eyigoz et al., 2020; Qiao et al., 2021). In addition, this step increases accuracy and stability of the predictions.

Linguistic features + Random Forest: Our second classification model is based on a Random Forest (RF) classifier[2] trained on various linguistic features extracted from the transcriptions. Our selection of linguistic features is motivated by previous research on dementia classification from spontaneous speech (Fraser et al., 2016; Balagopalan et al., 2020; Parsapoor et al., 2023; Liu et al., 2021; Syed et al., 2021a; Priyadarshinee et al., 2023; Eyigoz et al., 2020; Diaz-Asper et al., 2022; Tang et al., 2023). We implemented all linguistic features used in these approaches that were either a) reported as being important according to statistical

[1] Using HuggingFace's implementation in the `transformers` library v4.28 https://huggingface.co/bert-base-uncased.

[2] Using `scikit-learn`'s implementation v1.2.2, with 500 estimators and the default settings.

tests or feature importance analyses, or b) used by at least two studies. Our selection includes 15 syntactic features based on part-of-speech (POS) tags, 14 syntactic features based on grammatical constituents, 9 lexical features, along with 2 features of repetitiveness. A detailed list of all features and their definitions is given in Appendix B.

This RF-based approach has the significant advantage of providing explainability by quantifying feature contributions to individual predictions using SHAP values (Lundberg et al., 2020). This is in contrast to BERT, where explainability remains difficult.

3.4. Evaluation

Word Error Rate: We compare the ASR transcriptions to their manual counterparts using the Word Error Rate (WER), which is the most common evaluation metric for automatic speech recognition and defined as (Morris et al., 2004):

$$WER = \frac{S + D + I}{N} \quad (1)$$

where S, D, and I denote the number of word substitutions, deletions, and insertions, and N refers to the number of words in the reference transcription. We compare WER between ASR systems and between diagnosis (AD or control), age, and gender, using a generalized linear mixed-effect model[3] with WER as a dependent variable, age, gender, and ASR as fixed effects and subject as the random effect, resulting in the following formula (Wilkinson notation, Wilkinson and Rogers, 1973):

$$WER \sim ASR + label + age + gender + (1|subj) \quad (2)$$

Stability of features: We compute linguistic features for manual and ASR transcriptions. Let t_i^{man} be the manual transcription for sample $i \in \{1, ..., 156\}$, $t_i^{ASR_k}$ be the transcription from ASR system $k \in \{1, 2, 3\}$ for sample i, $f_j(t)$ be the value of feature j on transcript t. For each feature j and each sample i, we then compute the *relative difference* d between the feature calculated on the manual and ASR transcription as the absolute difference normalized by feature value:

$$d_{j,i,k} = \frac{f_j(t_i^{man}) - f_j(t_i^{ASR_k})}{\left(f_j(t_i^{man}) + f_j(t_i^{ASR_k})\right)/2} \quad (3)$$

We then estimate the *stability of a feature* j as the relative difference across all ASR systems k and samples i:

$$s_j = \frac{1}{3} \sum_k \left(\frac{1}{156} \sum_i |d_{j,i,k}| \right) \quad (4)$$

A value s_j close to zero indicates a *stable feature* j, having similar values for ASR and manual transcriptions. A large value s_j indicates an *unstable feature* j, with values changing strongly when we replace manual with ASR transcriptions.

End-to-end classification performance: To evaluate the end-to-end performance of the two machine-learning approaches on manual and ASR transcriptions, we use 10-fold cross validation, with identical random splits for all settings. We report the area under the ROC curve (AUROC) and binary accuracy for each setting. The AUROC is a popular metric for discriminative ability of a predictive model (Janssens and Martens, 2020), and does not require calibrated predictions nor the definition of a classification threshold. Since we do not focus on calibration, we use AUROC as our main metric. Results can vary when training the same setting multiple times, due to randomness in the initialization and training process. To capture this variability, we train and evaluate each setting 10 times, leading to a total of 100 training runs per setting[4], and we report mean and standard deviation for accuracy and AUROC.

We statistically analyze performance differences between settings by performing a series of pairwise permutation tests. A permutation test repeatedly shuffles AUROCs of both settings into two groups and computes the groups' difference in mean, then compares the observed difference to the randomly generated differences. For both BERT and RF, we compare the use of manual vs. ASR transcripts (2 tests), and we conduct pairwise comparisons between the different ASR systems (2×3 tests). In addition, we separately compare BERT against RF for manual and ASR transcripts (2 tests). This results in 10 tests, and we apply the Bonferroni Correction (Bonferroni, 1936) to counteract the multiple testing problem, resulting in a significance level of $0.05/10 = 0.005$.

Feature importance: The stability of a feature might influence its usefulness for AD classification given manual vs. ASR transcriptions. To evaluate this, we estimate feature importance, separately for manual transcriptions and for each ASR system. We quantify each feature's importance as the mean absolute SHAP value (Lundberg et al., 2020) in our RF-based approach.

[3] All statistical analyses were conducted using Python with a level of significance set at 0.05. The `statsmodel` library was used for the generalized linear mixed-effect model.

[4] For our BERT-based settings, we train 800 models, as 8 models are needed for each split (cf. Section 3.3).

4. Results

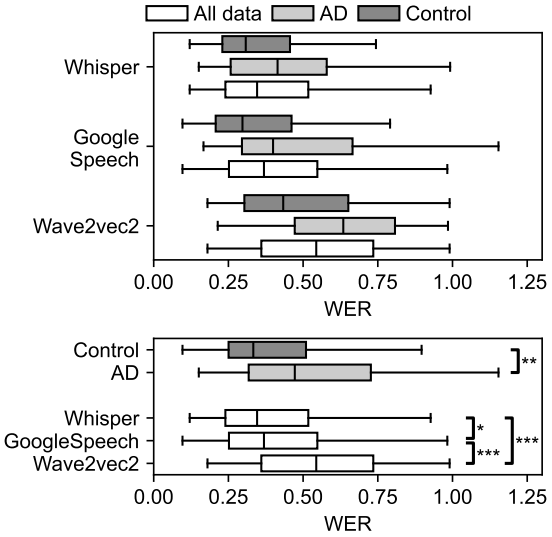

Figure 1: Distribution of Word Error Rate (WER) on the participant speech sections of ADReSS audio recordings. **Top:** For each ASR system, we show the distribution of all samples (white) and subgroups of AD patients and healthy controls (grey). **Bottom:** We show overall WER distribution for AD and control across ASR systems, as well as overall WER for each ASR system. Stars indicate statistical significance levels with $*$ ($p < 0.05$), $**$ ($p < 0.01$), and $***$ ($p < 0.001$).

Word Error Rate (WER): We observe that, across all ASR systems, WER differs strongly across subjects, ranging from almost perfect transcriptions to missing large fractions of spoken words. The statistical analysis reveals that Whisper (median WER: 0.35) performs significantly better than Google Speech (median WER: 0.37), which in turn significantly outperforms Wave2vec2 (median WER: 0.54). In addition, we observe a significant effect of diagnosis, showing that WER is lower (i.e. better) for healthy controls than for AD patients, while no significant effect is found for age and gender. Figure 1 shows the distribution of WER for each ASR, across all samples and separately for AD and control group. Detailed results of the statistical tests are given in Appendix C.

Stability of linguistic features: Figure 2 (A) shows the distribution of $d_{j,i,k}$, the relative difference between the features calculated on the manual and ASR transcriptions. Features are sorted according to their stability s_j across all ASR systems, with s_j explicitly displayed in subfigure (B). The most unstable features are `flesch_kincaid` (the Flesch–Kincaid grade level (Kincaid, 1975)), `avg_distance_between_utterances` (a feature of repetitiveness between sentences), `ROOT` $\rightarrow$ `FRAG` (the count of sentence fragments), and `words_not_in_dict_ratio` (the ratio of words not present in a dictionary of English words). The most stable features are `avg_word_length` (the average length of a word), as well as `mattr` and `brunets_index` (the moving-average type-token-ratio (Covington and McFall, 2010) and Brunét's index (Brunet et al., 1978), two measures of lexical richness).

Figure 2 (C) displays SHAP feature importance values for each feature (additional feature importance results are given in Appendix E). We observe that there is no clear trend in which important features are more or less stable than unimportant features, and feature importance often does not change dramatically when replacing manual with ASR transcriptions, even for very unstable features such as `flesch_kincaid`, which remains rather informative. This shows that, while feature values might change dramatically, they often do so for all participants, thereby retaining their discriminative power.

Approach	AUROC	Accuracy
Manual BERT	$\mathbf{0.899} \pm 0.009$	0.837 ± 0.013
Manual Lingu+RF	0.888 ± 0.003	0.821 ± 0.011
Google BERT	$\mathbf{0.837} \pm 0.010$	0.752 ± 0.015
Whisper BERT	0.801 ± 0.009	0.756 ± 0.015
Wave2vec2 BERT	0.819 ± 0.013	0.747 ± 0.013
Google Lingu+RF	$\mathbf{0.865} \pm 0.004$	0.792 ± 0.007
Whisper Lingu+RF	0.848 ± 0.004	0.785 ± 0.011
Wave2vec2 Lingu+RF	$\mathbf{0.860} \pm 0.004$	0.773 ± 0.006

Table 2: Accuracy and AUROC classification results (mean $\pm$ standard deviation over 10 runs) for different transcriptions and AD classification approaches. *Manual* refers to manual transcriptions, *Google*, *Whisper*, and *Wave2vec2* refer to ASR transcriptions. *BERT* represents the fine-tuned BERT approach, while *Lingu+RF* refers to the approach using linguistic features and Random Forest.

Classification performance for AD recognition: Table 2 presents results for all combinations of transcriptions and classification approaches. We report AUROC as our main metric, but also include accuracy results, for comparability with prior research. We make the following observations (detailed statistical results are given in Appendix C):

- *Manual transcriptions lead to better performance than ASR.* This is statistically significant and true for both BERT and RF.

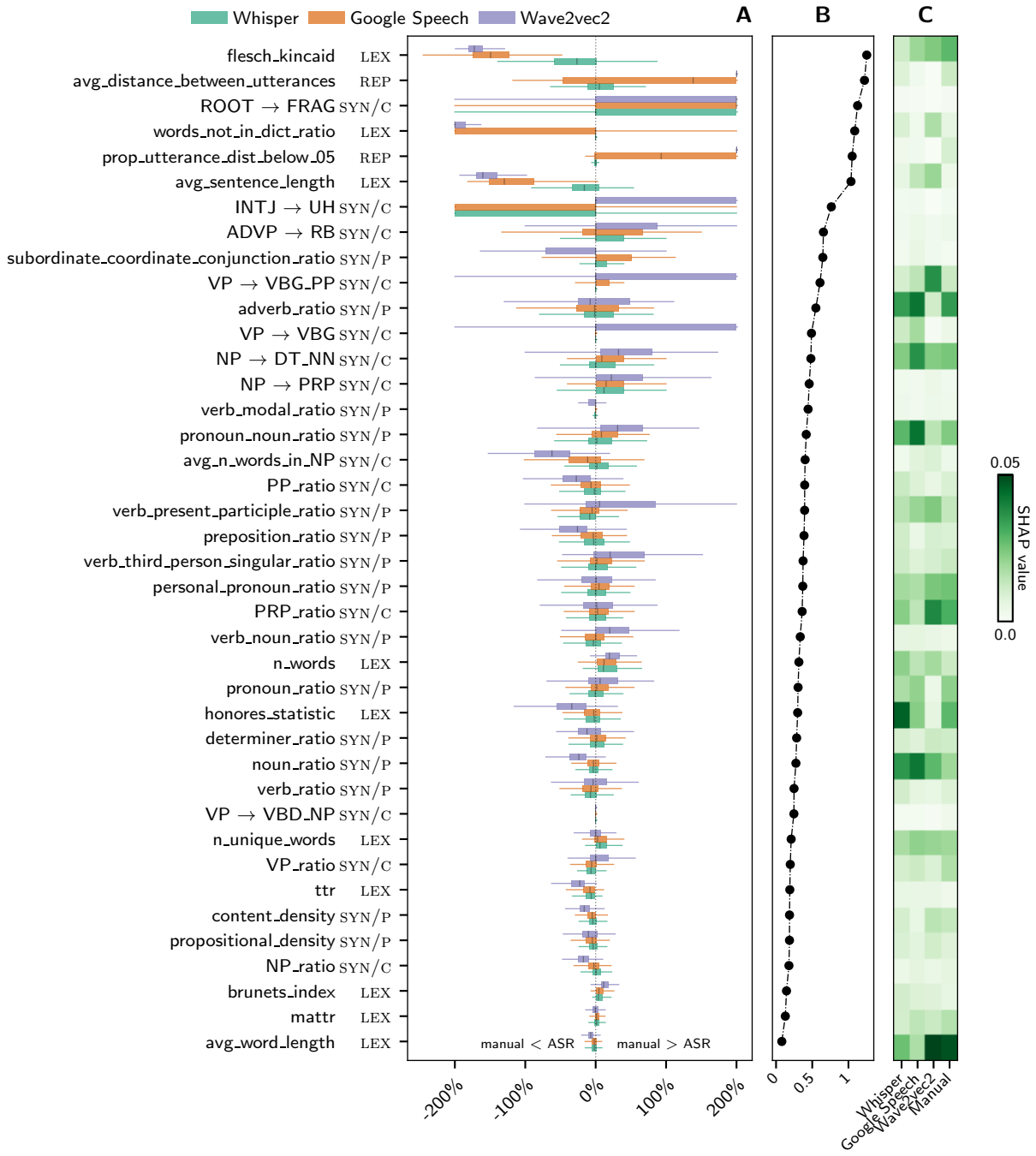

Figure 2: **A**: Distribution of the relative difference $d_{j,i,k}$ between manual and ASR transcripts, for each feature and each ASR system. Features are sorted according to their stability s_j (subfigure B). We display the feature group for each feature (syntactic features based on POS tags `SYN/P`, syntactic features based on grammatical constituents `SYN/C`, lexical features `LEX`, and features of repetitiveness `REP`). *Example:* The most unstable feature is the lexical (`LEX`) feature `flesch_kincaid`, the Flesch–Kincaid grade level (Kincaid, 1975), which is a combination of the number of words in a sentence and the number of syllables in a word. All ASR transcripts produce higher values of this feature than manual transcripts. This is caused by longer sentences produces by ASR compared to manual transcriptions. **B**: s_j, the stability of a feature across all ASR systems. Lower means more stable. **C**: Feature importance given by mean absolute SHAP value, based on the trained RF classifier, for manual transcripts and for each ASR. Dark-green indicates high feature importance, white indicates low feature importance.

- *When using ASR transcripts, linguistic features + Random Forest produces better results than fine-tuned BERT.* Fine-tuning a BERT model on manual transcriptions works better than the linguistic feature approach, a finding that confirms results reported by Balagopalan et al. (2020). However, when replacing manual with ASR transcripts, our BERT-based approach degrades more strongly, losing its advantage over and performing significantly worse than simpler and more interpretable linguistic features. This shows that the choice of the best algorithm depends on the use of ASR, and a simpler approach proves more robust to transcription errors.

- *Google Speech is the best overall ASR system w.r.t. AD classification performance.* It significantly outperforms Whisper and Wave2vec2 when using BERT. In the RF-based approach, Google Speech is significantly better than Whisper, while the difference to Wave2vec2 is not significant.

- *Lower (i.e. better) WER can lead to worse classification performance.* Whisper, the winner of the direct comparison of ASR systems w.r.t. WER, produces significantly lower AD classification performance than Google Speech and Wave2vec2.

5. Discussion and Conclusion

Word Error Rate: Our results show significant differences in transcription error rates (Whisper $<$ Google Speech $<$ Wave2vec2) and a consistent bias of higher error rates for AD speech than healthy controls. This is presumably a consequence of ASR's training on predominantly healthy speech, and presents a potential problem if such systems are used in a clinical setting. In addition, we have observed that WER differs strongly across subjects.

We believe that high WER and large variability across subjects are mainly caused by poor audio quality. The data[5] was originally collected in the 1980s, and manual inspection of the audio files shows quality issues such as significant background noise and faint voices caused by participants being far away from the microphone. In addition, we have observed that the noise reduction preprocessing applied by the dataset's authors (Luz et al., 2020) produces damaging side effects on the recordings, including partial removal of participant speech. Although this dataset remains critical for research on speech-based AD biomarkers, as it is the only balanced publicly available resource of AD spontaneous speech, better-quality recordings are important for the success of future research.

Our results also show that WER is lower (better) for healthy controls than for AD patients. This is consistent with prior work by Li et al. (2022); Soroski et al. (2022). We suspect that AD symptoms such as dysarthria (Cummings, 2020), resulting in unclear pronunciation, might be the underlying cause. However, Figure 1 indicates that WER variability is much larger between subjects within a group than between AD and control.

Stability of linguistic features: The influence of ASR on linguistic features varies between features. Some features change strongly when replacing manual with ASR transcriptions, while others remain rather stable. The instability of the most unstable features are caused by the following effects:

1. *Sentence boundaries:* Sentence boundaries in spoken language are often unclear. Wave2vec2 does not produce any boundaries (one long sentence), Google Speech and Whisper both produce longer sentences than utterances present in the manual transcription. This strongly affects i) `flesch_kincaid`, the Flesch–Kincaid grade level (Kincaid, 1975), ii) repetitiveness metrics between pairs of sentences (`prop_utterance_dist_below_05`, `avg_distance_between_utterances`), and iii) `avg_sentence_length`, the average number of words in a sentence.

2. *Out-of-dictionary words:* Wave2vec2 does not use a language model for decoding, and thus produces many misspelled words. Whisper and Google Speech produce some special tokens such as *mm-hmm*, "…", or *mhm*. `words_not_in_dict_ratio` is based on a dictionary of English words that does not contain these, and is thus very sensitive to such ASR-specific behaviours.

3. *Constituency parsing:* The syntactic features based on grammatical constituents `SYN/C` are based on linguistic constituency parsing. Small transcription errors can cause large changes in constructed parse trees, resulting in strong deviations in these features.

The most stable features are the average length of a word (`avg_word_length`), as well as syntactic features based on POS `SYN/P` and constituency ratios that only rely on local contexts (e.g. `NP_ratio`, the ratio of noun phrases).

[5]The ADReSS dataset is a subset of the larger but unbalanced DementiaBank English PITT corpus (Becker et al., 1994).

The usefulness of a feature in discriminating AD is not affected by its instability. Important features are not more or less stable than unimportant ones, and feature importance does not change systematically when replacing manual with ASR transcriptions. Although this allows results from studies using manual transcriptions to be generalized to the more naturalistic ASR setting, we recommend using caution when interpreting unstable features extracted from ASR. For example, conclusions on the nature of speech in AD should not be drawn from features based on sentence boundaries.

Classification performance for AD recognition: Classification performance degrades when using ASR instead of manual transcriptions in an end-to-end AD classification pipeline. This confirms Soroski et al. (2022)'s finding, but provides a stronger result than the more unclear picture reported by Li et al. (2021), Li et al. (2022), and Wang et al. (2022), where performance was comparable. Research on larger and more diverse datasets of AD spontaneous speech are needed to answer this question conclusively. Among the evaluated ASR systems, the Google Speech "Chirp" model performs best overall.

Importantly, when using ASR transcriptions, the BERT-based approach degrades much more strongly than the simpler linguistic feature-based approach, resulting in an inferior classification performance. We hypothesize that BERT's loss in performance is due to ASR transcription errors resulting in shorter, less coherent texts, reducing the usefulness of BERT's pre-training. Moreover, we conducted some additional experiments not presented here (results are given in Appendix D), where we fine-tuned a BERT model on the *entire audio*, including interviewer interventions. This improves BERT's performance, indicating that the additional context of the interviewer interactions is being picked up by the model. We consider this an undesirable behavior, as a potential future real-time detection system will likely be restricted to participant speech only. The influence of the interviewer on AD classification has also been discussed by Pérez-Toro et al. (2021), and it raises some concerns about the generalizability of prior approaches using the ADReSS dataset, as these presumably base at least part of their diagnostic power on these interviewer sections.

The RF-based approach has the advantage of providing explainability in the form of e.g. SHAP values (cf. Figure 2 (C)), in addition to its superior performance when faced with ASR transcriptions. Explainability of AI is essential in medical applications because it promotes trust (for patients and clinicians), transparency, and accountability, and addresses legal regulations (e.g. the EU's "AI Act"). Approaches of explainability for BERT exist, e.g. assigning word contributions using methods such as LIME (Ribeiro et al., 2016), Captum (Kokhlikyan et al., 2020), or TransSHAP (Kokalj et al., 2021), or analyzing the model's internal attention weights, using e.g. *BertViz* (Vig, 2019). However, these methods are harder to interpret, more difficult to communicate to patients and clinicians, and unable to evaluate the contribution of known AD symptoms or potential confounding factors. As an example, consider the length of a transcription, which is significantly shorter for AD than control (cf. Table 1). This is captured explicitly by our linguistic feature `n_words`, allowing feature importance analyses to quantify its contribution to classification performance. We are not aware of an easy way to examine this in BERT.

The relative robustness of our RF-based approach, combined with its additional explainability, serves as a reminder to continue research on established methods motivated by prior knowledge.

The effect of WER on AD classification performance: Surprisingly, a lower (better) WER does not translate into better AD classification performance in our experiments. This finding aligns with prior research (Tang et al., 2023; Li et al., 2022) and suggests that the WER metric is insufficient to capture the transcriptions' effectiveness for AD detection. For instance, despite Wave2vec2 producing more misspelled words, it shows superior end-to-end performance compared to Whisper, which has a lower WER. We hypothesize that the language model used in Whisper's decoding algorithm smooths the output, by e.g. removing hesitation markers, word repetitions, out-of-context expressions, or unintelligible terms, thereby losing information that might be relevant to discriminate between AD and healthy controls. In contrast, Wave2vec2 does not rely on a language model for decoding, resulting in more word-level errors, yet retaining valuable information contained in the recording, which renders it more useful for classification. Future research could develop more appropriate ASR quality metrics by imposing stronger penalties for missing words.

Limitations The main limitation of our study lies in the low number of samples in the dataset and the quality of the recordings. While we believe most of the observed trends will generalize to other data, our results are limited to the diversity of this dataset. Future research on other datasets could confirm and strengthen our findings.

Conclusion In conclusion, we have compared three popular ASR systems and recommend the recent Google Speech "Chirp" model, as it has

low WER and leads to high AD classification performance. Our experiments confirm the potential of ASR in a fully automatic AD detection system based on spontaneous speech. Popular linguistic features and Random Forest are more robust to ASR transcription errors than a fine-tuned BERT model. Despite a loss of performance compared to using manual transcriptions, classification AUROC of 0.87 remains strong. In addition, our results contribute to a better understanding of well-established linguistic features, assessing their stability when replacing manual with ASR transcriptions. Future research should investigate the generalizability of AD classification to other datasets. In addition, we plan on evaluating the fine-tuned BERT model in more detail, assessing the influence of transcription length and interviewer sections.

6. Acknowledgements

This work was supported by the University Research Priority Program (URPP) "Dynamics of Healthy Aging", and the University platform "Linguistic Research Infrastructure (LiRI)", Zurich, Switzerland.

7. Bibliographical References

Zoe Arvanitakis, Raj C Shah, and David A Bennett. 2019. Diagnosis and management of dementia. *Jama*, 322(16):1589–1599.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. 2020. To bert or not to bert: comparing speech and language-based approaches for alzheimer's disease detection. *arXiv preprint arXiv:2008.01551*.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. 1994. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of neurology*, 51(6):585–594.

Carlo Bonferroni. 1936. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62.

Étienne Brunet et al. 1978. *Le vocabulaire de Jean Giraudoux structure et évolution*. Slatkine.

Laura Calzà, Gloria Gagliardi, Rema Rossini Favretti, and Fabio Tamburini. 2021. Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. *Computer Speech & Language*, 65:101113.

Michael A Covington and Joe D McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100.

Louise Cummings. 2020. *Alzheimer's Dementia*, page 1–19. Cambridge University Press.

Nicholas Cummins, Yilin Pan, Zhao Ren, Julian Fritsch, Venkata Srikanth Nallanthighal, Heidi Christensen, Daniel Blackburn, Björn W Schuller, Mathew Magimai-Doss, Helmer Strik, et al. 2020. A comparison of acoustic and linguistics methodologies for alzheimer's dementia recognition. In *Interspeech 2020*, pages 2182–2186. ISCA-International Speech Communication Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Catherine Diaz-Asper, Chelsea Chandler, Raymond S Turner, Brigid Reynolds, and Brita Elvevåg. 2022. Increasing access to cognitive screening in the elderly: Applying natural language processing methods to speech collected over the telephone. *Cortex*, 156:26–38.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Elif Eyigoz, Sachin Mathur, Mar Santamaria, Guillermo Cecchi, and Melissa Naylor. 2020. Linguistic markers predict onset of alzheimer's disease. *EClinicalMedicine*, 28.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. 2016. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. 2022. Esb: A benchmark for multi-domain end-to-end speech recognition.

Harold Goodglass, Edith Kaplan, and Sandra Weintraub. 2001. *BDAE: The Boston diagnostic aphasia examination*. Lippincott Williams & Wilkins Philadelphia, PA.

Tony Honoré. 1979. Some simple measures of richness of vocabulary. 2010.

Loukas Ilias, Dimitris Askounis, and John Psarras. 2023. Detecting dementia from speech and transcripts using transformers. *Computer Speech & Language*, 79:101485.

A Cecile J W Janssens and Forike K Martens. 2020. Reflection on modern methods: Revisiting the area under the ROC Curve. *International Journal of Epidemiology*, 49(4):1397–1403.

J.P. Kincaid. 1975. *Derivation of New Readability Formulas: (automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Research Branch report. Chief of Naval Technical Training, Naval Air Station Memphis.

Enja Kokalj, Blaž Škrlj, Nada Lavrač, Senja Pollak, and Marko Robnik-Šikonja. 2021. BERT meets shapley: Extending SHAP explanations to transformer-based classifiers. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 16–21, Online. Association for Computational Linguistics.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.

Lampros C Kourtis, Oliver B Regele, Justin M Wright, and Graham B Jones. 2019. Digital biomarkers for alzheimer's disease: the mobile/wearable devices opportunity. *NPJ digital medicine*, 2(1):9.

Changye Li, Trevor Cohen, and Serguei Pakhomov. 2022. The far side of failure: Investigating the impact of speech recognition errors on subsequent dementia classification.

Jinchao Li, Jianwei Yu, Zi Ye, Simon Wong, Manwai Mak, Brian Mak, Xunying Liu, and Helen Meng. 2021. A comparative study of acoustic and linguistic features classification for alzheimer's disease detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6423–6427. IEEE.

Ziming Liu, Lauren Proctor, Parker N Collier, and Xiaopeng Zhao. 2021. Automatic diagnosis and prediction of cognitive decline associated with alzheimer's dementia through spontaneous speech. In *2021 ieee international conference on signal and image processing applications (icsipa)*, pages 39–43. IEEE.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839.

Saturnino Luz, Sofia de la Fuente, and Pierre Albert. 2018. A method for analysis of patient speech in dialogue for dementia detection. *arXiv preprint arXiv:1811.09919*.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge. In *Proceedings of INTERSPEECH 2020*, Shanghai, China.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2021. Detecting Cognitive Decline Using Speech Only: The ADReSSo Challenge. In *Proc. Interspeech 2021*, pages 3780–3784.

Brian MacWhinney. 2000. The childes project: tools for analyzing talk. *Child Language Teaching and Therapy*, 8.

Matej Martinc, Fasih Haider, Senja Pollak, and Saturnino Luz. 2021. Temporal integration of text transcripts and acoustic features for alzheimer's diagnosis based on spontaneous speech. *Frontiers in Aging Neuroscience*, 13:642647.

Vaden Masrani, Gabriel Murray, Thalia Field, and Giuseppe Carenini. 2017. Detecting dementia through retrospective analysis of routine blog posts by bloggers with dementia. In *BioNLP 2017*, pages 232–237.

Jon F. Miller. 1981. Assessing language production in children: Experimental procedures.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Eighth International Conference on Spoken Language Processing*.

Yilin Pan, Bahman Mirheidari, Jennifer M Harris, Jennifer C Thompson, Matthew Jones, Julie S Snowden, Daniel Blackburn, and Heidi Christensen. 2021. Using the outputs of different automatic speech recognition paradigms for acoustic-and bert-based alzheimer's dementia detection through spontaneous speech. In *Interspeech*, pages 3810–3814.

Mahboobeh Parsapoor, Muhammad Raisul Alam, and Alex Mihailidis. 2023. Performance of machine learning algorithms for dementia assessment: impacts of language tasks, recording media, and modalities. *BMC Medical Informatics and Decision Making*, 23(1):45.

Paula Andrea Pérez-Toro, Sebastian P Bayerl, Tomas Arias-Vergara, Juan Camilo Vásquez-Correa, Philipp Klumpp, Maria Schuster, Elmar Nöth, Juan Rafael Orozco-Arroyave, and Korbinian Riedhammer. 2021. Influence of the interviewer on the automatic assessment of alzheimer's disease in the context of the adresso challenge. In *Interspeech*, pages 3785–3789.

Prachee Priyadarshinee, Christopher Johann Clarke, Jan Melechovsky, Cindy Ming Ying Lin, Balamurali BT, and Jer-Ming Chen. 2023. Alzheimer's dementia speech (audio vs. text): Multi-modal machine learning at high vs. low resolution. *Applied Sciences*, 13(7):4244.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Yu Qiao, Xuefeng Yin, Daniel Wiechmann, and Elma Kerz. 2021. Alzheimer's Disease Detection from Spontaneous Speech Through Combining Linguistic Complexity and (Dis)Fluency Features with Pretrained Language Models. In *Proc. Interspeech 2021*, pages 3805–3809.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Alireza Roshanzamir, Hamid Aghajan, and Mahdieh Soleymani Baghshah. 2021. Transformer-based deep neural network language models for alzheimer's disease risk assessment from targeted speech. *BMC Medical Informatics and Decision Making*, 21:1–14.

Thomas Soroski, Thiago da Cunha Vasco, Sally Newton-Mason, Saffrin Granby, Caitlin Lewis, Anuj Harisinghani, Matteo Rizzo, Cristina Conati, Gabriel Murray, Giuseppe Carenini, et al. 2022. Evaluating web-based automatic transcription for alzheimer speech data: Transcript comparison and machine learning analysis. *JMIR aging*, 5(3):e33460.

Zafi Sherhan Syed, Muhammad Shehram Shah Syed, Margaret Lech, and Elena Pirogova. 2021a. Automated recognition of alzheimer's dementia using bag-of-deep-features and model ensembling. *IEEE Access*, 9:88377–88390.

Zafi Sherhan Syed, Muhammad Shehram Shah Syed, Margaret Lech, and Elena Pirogova. 2021b. Tackling the adresso challenge 2021: The muet-rmit system for alzheimer's dementia recognition from spontaneous speech. In *Interspeech*, pages 3815–3819.

Lijuan Tang, Zhenglin Zhang, Feifan Feng, Li-Zhuang Yang, and Hai Li. 2023. Explainable alzheimer's disease detection using linguistic features from automatic speech recognition. *Dementia and Geriatric Cognitive Disorders*.

Malgorzata Anna Ulasik, Manuela Hürlimann, Fabian Germann, Esin Gedik, Fernando Benites, and Mark Cieliebak. 2020. CEASR: A corpus for evaluating automatic speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6477–6485, Marseille, France. European Language Resources Association.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Yi Wang, Tianzi Wang, Zi Ye, Lingwei Meng, Shoukang Hu, Xixin Wu, Xunying Liu, and Helen Meng. 2022. Exploring linguistic feature and model combination for speech recognition based automatic AD detection. In *Proc. Interspeech 2022*, pages 3328–3332.

Jochen Weiner, Mathis Engelbart, and Tanja Schultz. 2017. Manual and Automatic Transcriptions in Dementia Detection from Speech. In *Proc. Interspeech 2017*, pages 3117–3121.

GN Wilkinson and CE Rogers. 1973. Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 22(3):392–399.

Bengt Winblad, Anders Wimo, and Linus Jönsson. 2006. O1–05–08: The worldwide directs

costs and costs of informal care of dementia. *Alzheimer's & Dementia*, 2:S19–S20.

Jiahong Yuan, Yuchen Bian, Xingyu Cai, Jiaji Huang, Zheng Ye, and Kenneth Church. 2020. Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease. In *Interspeech*, volume 2020, pages 2162–6.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2020. Revisiting few-sample bert fine-tuning. *arXiv preprint arXiv:2006.05987*.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*.

Appendix A: Preprocessing steps for CHAT transcriptions

The manual of the CHAT transcription format (MacWhinney, 2000) is available online[6]. We only retain main lines of participant utterances, ignoring any dependent tier lines. We apply the following preprocessing steps to each utterance. Our implementation is available on Github[7]:

1. Remove events such as `&=clears:throat`
2. Remove complex local events of the form `[^ text]`
3. Remove special sign `&` for e.g. disfluencies, but keep phonological fragment, fillers, nonwords, interposed words (`&um`, etc.)
4. Remove omitted words (`0`)
5. Remove unintelligible speech sign `xxx`
6. Remove sign for letter transcription (e.g. make `m` out of `m@l`)
7. Remove replacement notation (annotator marking what is meant by a participant word) (e.g. `chair [: stool]` → `chair`
8. Remove pauses (`(.)`, `(..)`, `(...)`)
9. Remove special marking of non-complete words (omitted parts), keeping the full word
10. Remove sign for retracing, reformulation, false start without retracing, unclear retracing etc. (`[//]`, `[///]` etc.), but keep utterance
11. Remove custom postcodes (e.g. `[+ jar]`)
12. Remove error markings (e.g. `[* s:uk]`)
13. Make word repetitions explicit: `get [x 3]` → `get get get`
14. Remove special utterances terminators, like trailing off
15. Remove overlap markers `[>]`, `[< ]`, `+<`
16. Remove overlap signs `<text>`
17. Remove unibet transcription words (e.g. `k t ə @u`)
18. Remove quotation marks
19. Remove interruption signs (`+,`, `+/.` `+/?` `+//.` `+//?`
20. Remove transcription break (`+.`, self completion `+,` other completion `++`, and quick uptake `+^`)
21. Normalize linkages and *irregular combinations* (e.g. `kind_of`, `how_about`)
22. Remove Lengthened Syllable marker (e.g. `s:tool`)
23. Replace satellite markers ‡ and „ by comma
24. Remove special form markers (e.g. `xxx@a`, `bingo@o`)
25. Remove compound marker `+` (e.g. `bird+house` → `birdhouse`)

Appendix B: Linguistic features

A detailed list of the linguistic features is presented in Table 3. These features are motivated by previous research on dementia classification from spontaneous speech (cf. Section 3.3):

(1) Fraser et al. (2016)
(2) Parsapoor et al. (2023)
(3) Liu et al. (2021)
(4) Syed et al. (2021a)
(5) Priyadarshinee et al. (2023)
(6) Eyigoz et al. (2020)
(7) Balagopalan et al. (2020)
(8) Diaz-Asper et al. (2022)
(9) Tang et al. (2023)

In addition to the inclusion criteria presented in Section 3.3, we excluded features that

- Were not described precisely enough to allow reimplementation.
- Had a constant feature value for all participants. Namely, these include features based on grammatical constituents that did not appear in our dataset.

We use the Stanza NLP library (Qi et al., 2020) for POS and constituency parsing[8]. The code of our implementation can be accessed on GitHub[9].

[6] https://talkbank.org/manuals/CHAT.html
[7] https://github.com/jheitz/lrec_coling2024_asr_paper/blob/main/src/dataloader/chat_parser.py
[8] Version 1.5.0
[9] https://github.com/jheitz/lrec_coling2024_asr_paper/blob/main/src/preprocessing/linguistic_features_literature.py

Group	Feature Name	Description	Used by prior research
SYN/P	pronoun_noun_ratio	Ratio of pronouns to nouns	(1), (7), (3)
SYN/P	verb_noun_ratio	Ratio of verbs to nouns	(3)
SYN/P	subordinate_coordinate_conjunction_ratio	Ratio of subordinate to coordinate conjunctions	(2)
SYN/P	adverb_ratio	Ratio of adverbs to all words	(1), (7), (9)
SYN/P	noun_ratio	Ratio of nouns to all words	(1), (8), (9)
SYN/P	verb_ratio	Ratio of verbs to all words	(1), (9)
SYN/P	pronoun_ratio	Ratio of pronouns to all words	(7), (9)
SYN/P	personal_pronoun_ratio	Ratio of personal pronouns to all words	(7)
SYN/P	determiner_ratio	Ratio of determiners to all words	(8)
SYN/P	preposition_ratio	Ratio of prepositions to all words	(9)
SYN/P	verb_present_participle_ratio	Ratio of verb (present participle) to all words	(7), (8)
SYN/P	verb_modal_ratio	Ratio of modal verbs to all words	(8)
SYN/P	verb_third_person_singular_ratio	Ratio of verbs in 3rd person singular to all words	(1)
SYN/P	propositional_density	Based on POS tags, according to Parsapoor et al. (2023)	(2), (6)
SYN/P	content_density	Based on POS tags, cccording to Parsapoor et al. (2023)	(2), (8), (9)
SYN/C	NP→PRP	Count of CFG production rules acc. to constituency parsing	(1)
SYN/C	ADVP→RB	Count of CFG production rules acc. to constituency parsing	(1), (7)
SYN/C	NP→DT_NN	Count of CFG production rules acc. to constituency parsing	(1)
SYN/C	ROOT→FRAG	Count of CFG production rules acc. to constituency parsing	(1)
SYN/C	VP→AUX_VP	Count of CFG production rules acc. to constituency parsing	(1)
SYN/C	VP→VBG	Count of CFG production rules acc. to constituency parsing	(1)
SYN/C	VP→VBG_PP	Count of CFG production rules acc. to constituency parsing	(1)
SYN/C	VP→IN_S	Count of CFG production rules acc. to constituency parsing	(1)
SYN/C	VP→AUX_ADJP	Count of CFG production rules acc. to constituency parsing	(1)
SYN/C	VP→AUX	Count of CFG production rules acc. to constituency parsing	(1)
SYN/C	VP→VBD_NP	Count of CFG production rules acc. to constituency parsing	(1)
SYN/C	INTJ→UH	Count of CFG production rules acc. to constituency parsing	(1)
SYN/C	NP_ratio	Ratio to all constituents	(9)
SYN/C	PRP_ratio	Ratio to all constituents	(9)
SYN/C	PP_ratio	Ratio to all constituents	(1)
SYN/C	VP_ratio	Ratio to all constituents	(1)
SYN/C	avg_n_words_in_NP	Avg. number of words in noun phrase	(9)
LEX	flesch_kincaid	The Flesch–Kincaid Grade Level Formula (Kincaid, 1975), a metric of readability.	(2)
LEX	avg_word_length	Average letters per word	(1), (7)
LEX	n_words	Number of words in transcript	(9), (3), (5), (8)
LEX	n_unique_words	Number of unique words in transcript	(5), (8)
LEX	avg_sentence_length	Average number of words per sentence	(3)
LEX	words_not_in_dict_ratio	Ratio of words not in English dictionary (Used dictionary)	(1), (7)
LEX	brunets_index	Brunét's index (Brunet et al., 1978), a metric of lexical richness defined as $N^{V^{-0.165}}$, with N the number of words and V is the number of unique words	(2), (8)
LEX	honores_statistic	Honoré Statistic (Honoré, 1979), a metric of lexical richness defined as $\frac{100\log(N)}{1-V_1/V}$, with N the number of words and V is the number of unique words, and V_1 the number of unique words appearing once	(1), (9), (2), (8)
LEX	ttr	The type-token-ratio (TTR) (Miller, 1981), a measure of lexical diversity, defined as number of words devided by number of unique words.	(3), (8)
LEX	mattr	The moving-average type-token-ratio (Covington and McFall, 2010) with window length 20.	(8)
REP	avg_distance_between_utterances	Avg. cosine distance between utterances in transcript, a feature of repetitiveness, based on Masrani et al. (2017)'s implementation (Availabe on GitHub)	(1), (7)
REP	prop_utterance_dist_below_05	Proportion of sentence pairs where cosine distance $\leq$ 0.5, based on Masrani et al. (2017)'s implementation	(1), (7)

Table 3: Table of all used linguistic features: Feature groups are: Syntactic features based on POS tags SYN/P, syntactic features based on grammatical constituents SYN/C, lexical features LEX, and features of repetitiveness REP.

Appendix C: Statistical test results

Word Error Rate (WER)

Table 4 reports results of the generalized linear mixed-effect model (Formula 2).

Mixed Linear Model Regression Results

Model:	MixedLM	Dependent Variable:	WER
No. Observations:	468	Method:	REML
No. Groups:	156	Scale:	0.0104
Min. group size:	3	Log-Likelihood:	174.3157
Max. group size:	3	Converged:	Yes
Mean group size:	3.0		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	0.633	0.185	3.412	0.001	0.269	0.996
ASR[T.wave2vec2]	0.125	0.012	10.785	0.000	0.102	0.147
ASR[T.whisper]	−0.024	0.012	−2.090	0.037	−0.047	−0.002
gender[T.1]	0.012	0.037	0.320	0.749	−0.061	0.085
Label[T.1]	0.118	0.037	3.208	0.001	0.046	0.191
age	−0.004	0.003	−1.387	0.166	−0.009	0.002
Group Var	0.050	0.073				

Table 4: Generalized linear mixed-effect model to test for a statistical effects of ASR system, age, gender, and diagnosis (with $Label = 1$ (AD), $Label = 0$ (Control)) on Word Error Rate (WER).

End-to-end AD classification

Table 5 presents detailed results of the permutation tests comparing different AD classification settings.

Test	Observed	P-Value
Manual transcripts: BERT vs. Linguistic+RF:	0.0105	*0.0026
ASR transcripts: BERT vs. Linguistic+RF:	-0.0385	*0.0000
BERT: Manual vs. ASR transcripts	0.0794	*0.0000
Linguistic+RF: Manual vs. ASR transcripts	0.0305	*0.0000
BERT: GoogleSpeech vs. Wave2vec2	0.0178	*0.0030
BERT: GoogleSpeech vs. Whisper	0.0362	*0.0000
BERT: Wave2vec2 vs. Whisper	0.0184	*0.0047
Linguistic+RF: GoogleSpeech vs. Wave2vec2	0.0048	0.0196
Linguistic+RF: GoogleSpeech vs. Whisper	0.0171	*0.0000
Linguistic+RF: Wave2vec2 vs. Whisper	0.0123	*0.0000

Table 5: Permutation tests comparing different AD classification settings. For a test A vs. B, the *Observed* represents the difference between mean AUROC across 10 runs of the two settings. *P-Values* are calculated as the fraction of times the absolute value of *Observed* is smaller than the absolute values of the differences in means in the permuted distributions, with * indicating significant results according to the significance level $0.05/10 = 0.005$.

Appendix D: Additional results including interviewer interactions

Table 6 present classification results in two versions: a) removing the interviewer section from audio and transcriptions (as presented in the main paper), b) retaining the interviewer section in audio and transcript. We observe that the BERT-based approach profits from the inclusion of the interviewer sections, indicating that this additional context is being picked up by the model.

Appendix E: Feature importance analysis

Figure 3 displays an overview of feature importance for all linguistic features. This figure contains the same information as Figure 2 (C), but the ordering of features simplifies interpretation.

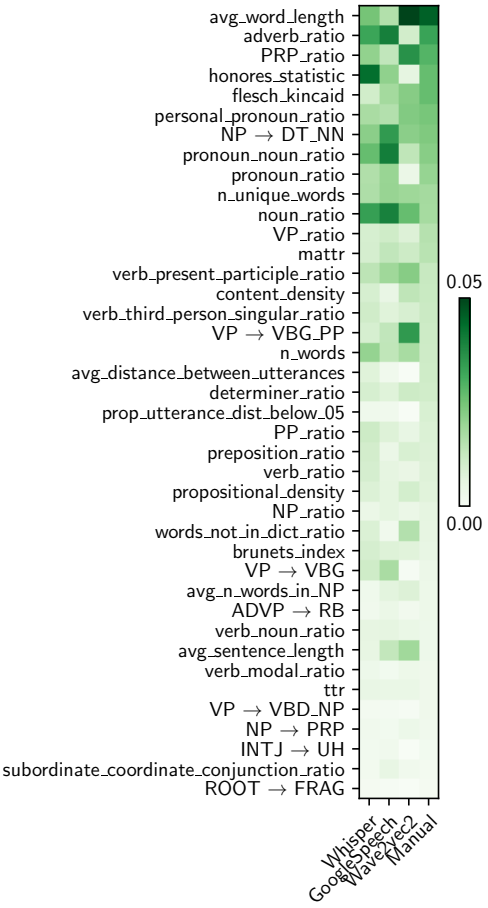

Figure 3: Mean absolute SHAP value as a metric of feature importance for all linguistic features. Features are sorted by their importance when used on manual transcripts (right-most column).

	Without Interviewer		With Interviewer	
Approach	AUROC	Accuracy	AUROC	Accuracy
Manual BERT	0.899 ± 0.009	0.837 ± 0.013	0.884 ± 0.012	0.797 ± 0.020
Manual Lingu+RF	0.888 ± 0.003	0.821 ± 0.011	0.886 ± 0.003	0.794 ± 0.007
Google BERT	0.837 ± 0.010	0.752 ± 0.015	$\mathbf{0.855} \pm 0.005$	0.790 ± 0.013
Whisper BERT	0.801 ± 0.009	0.756 ± 0.015	$\mathbf{0.827} \pm 0.005$	0.765 ± 0.011
Wave2vec2 BERT	0.819 ± 0.013	0.747 ± 0.013	$\mathbf{0.840} \pm 0.006$	0.778 ± 0.014
Google Lingu+RF	0.865 ± 0.004	0.792 ± 0.007	0.862 ± 0.003	0.787 ± 0.011
Whisper Lingu+RF	0.848 ± 0.004	0.785 ± 0.011	0.816 ± 0.005	0.758 ± 0.010
Wave2vec2 Lingu+RF	0.860 ± 0.004	0.773 ± 0.006	0.797 ± 0.004	0.733 ± 0.014

Table 6: Accuracy and AUROC classification results (mean $\pm$ standard deviation over 10 runs) for different transcriptions and AD classification approaches. *Manual* refers to manual transcriptions, *Google*, *Whisper*, and *Wave2vec2* refer to ASR transcriptions. *BERT* represents the fine-tuned BERT approach, while *Lingu+RF* refers to the approach using linguistic features and Random Forest. Results are presented excluding interviewer utterances (*Without Interviewer*, as described in the main paper) and with interviewer utterances (*With Interviewer*, additional results).