A. List of features

List of lexico-syntactic features is in Table A.1, acoustic features in Table A.2 and semantic in Table A.3, all with brief descriptions and counts of sub-types. Spacy [2] is used for part-of-speech tagging in the linguistic pipeline, where tags belong to Penn Treebank[3].

B. Hyper-parameter Settings for Fine-tuning BERT

All our experiments are based on the *bert-base-uncased* variant [11], which consists of 12 layers, each having a hidden size of 768 and 12 attention heads. Maximum input length is 512 tokens. Initial learning rate is set to $2e-5$, and Adam optimizer [27] is used. Cross-entropy loss is used while fine-tuning for AD detection.

While the base BERT model is pre-trained with sentence pairs, our input to the model consists of speech transcripts with several transcribed utterances with start and separator special tokens from the BERT vocabulary at the beginning and end of each utterance respectively, following Liu et al. [26]. This is performed to ensure that utterance boundaries are easily encoded, since cross-utterance information such as coherence and utterance transitions is important for reliable AD detection [6]. An embedding, following Devlin *et al.* [11], pooling information across all tokenized units in the transcript is extracted as the aggregate transcript representation from the BERT base for each transcript. This is then passed to the classification layer, and the combined model is fine-tuned on the AD detection task – all using an open-source PyTorch [28] implementation of BERT-based text sequence classification models and tokenizers [25]. As noted by Devlin *et al.* [11], this pooled embedding representation heavily depends on the fine-tuning task – in our case, AD detection at transcript level.

C. Hyper-parameter Settings for Feature Classification

Hyper-parameters were tuned using grid search with 10-fold cross validation on the ADReSS challenge 'train' set.

The random forest classifier fits 200 decision trees and considers $\sqrt{features}$ when looking for the best split. The minimum number of samples required to split an internal node is 2, and the minimum number of samples required to be at a leaf node is 2. Bootstrap samples are used when building trees. All other parameters are set to the default value.

The Gaussian Naive Bayes classifier is fit with balanced priors and variance smoothing coefficient set to $1e-10$ and all other parameters default in each case..

The SVM is trained with a radial basis function kernel with kernel coefficient(γ) 0.001, and regularization parameter set to 100.

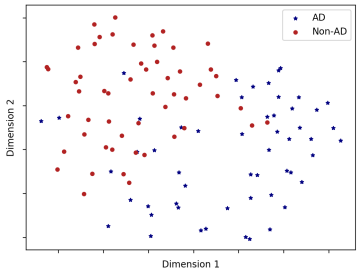
The NN used consists of 2 layers of 10 units each (note we varied both the number of units and number of layers while tuning for the optimal hyperparameter setting). The ReLU activation function is used at each hidden layer. The model is trained using Adam [27] for 200 epochs and with a batch size of number of samples in train set in each fold. All other parameters are default.

[2]https://spacy.io/usage/linguistic-features

[3]http://www.cis.upenn.edu/ treebank/

D. t-SNE Visualization

Figure 1: *A t-SNE plot showing class separation. Note we only use the 13 features significantly different between classes (see Table 2) in feature representation for this plot.*



In order to visualize the class-separability of the feature-based representations, we visualize t-SNE [35] plots in Figure 1. We observe strong class-separation between the two classes, indicating that a non-linear model would be capable of good AD detection performance with these representations.

E. Test Performance Metrics

The procedure for obtaining performance metrics on the test set was as follows:

1. Predictions from up to 5 models are sent to the challenge organizer for each prediction task – we sent predictions from 5 AD vs non-AD classification models (SVM, NN, RF, NB, BERT) and 5 regression models (linear and ridge regression).
2. Organizers send performance scores on the test set for each prediction set, which are then reported in Table 5 and Table 6.

Table A.1: *Summary of all lexico-syntactic features extracted. The number of features in each subtype is shown in the second column (titled "#features").*

Feature type	#Features	Brief Description
Syntactic Complexity	36	L2 Syntactic Complexity Analyzer [34] features; max/min utterance length, depth of syntactic parse tree
Production Rules	104	Number of times a production type occurs divided by total number of productions
Phrasal type ratios	13	Proportion, average length and rate of phrase types
Lexical norm-based	12	Average norms across all words, across nouns only and across verbs only for imageability, age of acquisition, familiarity and frequency (commonness)
Lexical richness	6	Type-token ratios (including moving window); brunet; Honoré's statistic
Word category	5	Proportion of demonstratives (e.g., "this"), function words, light verbs and inflected verbs, and propositions (POS tag verb, adjective, adverb, conjunction, or preposition)
Noun ratio	3	Ratios nouns:(nouns+verbs); nouns:verbs; pronouns:(nouns+pronouns)
Length measures	1	Average word length
Universal POS proportions	18	Proportions of Spacy univeral POS tags
POS tag proportions	53	Proportions of Penn Treebank POS tags
Local coherence	15	Avg/max/min similarity between word2vec [29] representations of utterances (with different dimensions)
Utterance distances	5	Fraction of pairs of utterances below a similarity threshold (0.5,0.3,0); avg/min distance
Speech-graph features	13	Representing words as nodes in a graph and computing density, number of loops, etc.
Utterance cohesion	1	Number of switches in verb tense across utterances divided by total number of utterances
Rate	2	Ratios – number of words: duration of audio; number of syllables: duration of speech,
Invalid words	1	Proportion of words not in the English dictionary
Sentiment norm-based	9	Average sentiment valence, arousal and dominance across all words, noun and verbs

Table A.2: *Summary of all acoustic features extracted. The number of features in each subtype is shown in the second column (titled "#features").*

Feature type	#Features	Brief Description
Pauses and fillers	9	Total and mean duration of pauses;long and short pause counts; pause to word ratio; fillers(um,uh); duration of pauses to word durations
Fundamental frequency	4	Avg/min/max/median fundamental frequency of audio
Duration-related	2	Duration of audio and spoken segment of audio
Zero-crossing rate	4	Avg/variance/skewness/kurtosis of zero-crossing rate
Mel-frequency Cepstral Coefficients (MFCC)	168	Avg/variance/skewness/kurtosis of 42 MFCC coefficients

Table A.3: *Summary of all semantic features extracted. The number of features in each subtype is shown in the second column (titled "#features").*

Feature type	#Features	Brief Description
Word frequency	10	Proportion of lemmatized words, relating to the Cookie Theft picture content units to total number of content units
Global coherence	15	Avg/min/max cosine distance between word2vec [29] utterances and picture content units, with varying dimensions of word2vec