

A Feature list

A.1 Token-Level features

For each token, we extract each of the features listed in Tab. 5. To ensure that each pause has at least one token preceding and succeeding it, *start* and *end* tokens are added to each utterance in DB (Sec. 3.1). For tokens that did not have these features, such as pauses and start/end tokens, all of the features were given a value of zero with the exception of the part of speech. For each feature, with the exception of word length and part of speech, we also extracted the same feature value from the lemmatized token. These features, along with a 5-dimensional, randomly initialized embedding for the part-of-speech, make up the 23-dimensions of each token input vector. Part-of-speech tagging is performed using Spacy (Honnibal and Montani, 2017). All word tokens then have missing values imputed with feature means, and are then normalized with respect to the feature means and standard deviations of the word tokens (i.e. excluding pauses, start/end tokens).

Identical subsequences found in both classes were removed from each data subset. Additionally, if multiple identical subsequences were found in only one of the classes, all but one of them are removed. Furthermore, we do not include utterances that include only a single pause as the final token, as we assume that this pause is occurring between consecutive sentences, rather than within a single sentence.

A.2 Transcript-Level Features

We classify transcripts using 500+ extracted features based on previous literature (Fraser et al., 2016; Tóth et al., 2018), which we refer to as the *Original* feature set. Each of these features come from one of 8 categories:

- **Information Units:** Semantic measures, pertaining to the ability to describe concepts and objects in the picture.
- **Discourse Mapping:** Features that help identify cohesion in speech using a visual representation of message organization in speech. We represent each word as a node to build a ‘speech graph’ (Mota et al., 2012), for the whole transcript. Examples of features extracted include number of edges in this graph, number of self-loops etc.

- **Coherence:** Semantic continuity that listeners perceive between utterances (locally or globally).
- **Lexical Complexity and Richness :** Different measures of lexical qualities and variation. Examples of features include average age of acquisition, number of occurrence of various POS tags etc.
- **Sentiment:** Sentiment lexical norms from Warriner *et al.* 2013. Examples include average sentiment valence over verbs, average sentiment dominance over nouns etc.
- **Syntactic Complexity:** Different measures to analyze the syntactic complexity of speech including features such as number of occurrence of various production rules, mean length of clause (in words) etc.
- **Word finding Difficulty:** Features quantifying difficulty in finding the right words. These include various pause features such as number of filled pauses, pause word ratio etc.
- **Acoustic:** Voice markers such as MFCC coefficients and Zero Crossing Rate (ZCR) related features.

All transcript-level features, including the aggregates described in Sec. A.3, have missing values imputed with feature medians. Features are then standardized by removing the feature median, and scaled according to the range between the first and third quartile (Pedregosa et al., 2011).

A.3 Transcript-Level Aggregates

In Sec. 4, we extend the *Original* set of features with transcript-level aggregates of token-level features. These include averages for each of the features described in Sec. A.1, with the exception of part of speech. Pauses are not considered when calculating the mean feature values. Additionally, for each part of speech (POS), we include the total amount of times that POS occurs at a certain distance from the pause, divided by the total number of pauses in the transcript multiplied by the percent of words in the transcript that are that POS.

B Classification

B.1 Subsequence Classification

In this work, we perform five-fold cross validation with each of the subsequence data subsets on sev-

Feature	# of features	Description
Word length	2	Length of the word, both in syllables and letters.
Sentiment	6	Three measures of the type and intensity of reaction a word produces.(Warriner et al., 2013)
Concreteness	2	Measure of the degree to which a word refers to a perceptible entity. (Brysbaert et al., 2014)
Imageability	2	How easy it is for a word to elicit a mental image. (Stadthagen-Gonzalez and Davis, 2006)
Age of acquisition	2	Average age that the word is learned. (Kuperman et al., 2012)
Frequency	2	Word counts in a corpus of over 385 million words. (Davies, 2009)
Familiarity	2	The perceived popularity of a word.(Stadthagen-Gonzalez and Davis, 2006)
Part of Speech	5	Grammatical category of the word.

Table 5: Token-level linguistic features used as input to the models during subsequence classification and their description.

Model	Bidirectional	# of Layers	Dropout	Epochs	Learning Rate	Momentum	λ	Batch Size	Approx. Time
M-C1	False(12)	2(10, 5)	False	600	0.01	0.9	0.0001	20	6 Min.
M-C2	True(12)	2(10, 5)	True(p=0.5)	600	0.01	0.9	0.0001	20	17 Min.
M-C3	False(50)	1(40)	True(p=0.5)	600	0.01	0.9	0.0001	20	27 Min.
M-Utt	False(50)	2(40, 20)	True(p=0.5)	600	0.01	0.9	0.0001	20	90 Min.

Table 6: Hyperparameters used by the best performing models for each data subset in the subsequence classification task, as well as the approximate time required to complete cross validation. The number of hidden units in the GRU is indicated in the ‘‘Bidirectional’’ column, and the number of hidden units in each layer of the predicting network is indicated in the ‘‘# of Layers’’ column.

Feature Set	Model	# of Features Selected	SMOTE
Original	Ens	-	False
Original w/ feat.sel	Ens	85	False
Original + F-D1	NN	11	False
Original + F-D2	Ens	15	False
Original + F-D3	Ens	5	True
Original + F-C2	NN	20	False
Original + F-C3	Ens	5	True

Table 7: Parameters used by the best performing models for each feature set in the transcript classification task.

eral different GRU based models (Cho et al., 2014) with attention (Yang et al., 2016). Each model consists of a GRU that takes the subsequences as input, and outputs to a feed forward neural network which then makes predictions. The attention mechanism uses a linear layer that has as many hidden units as the GRU, as well as a context vector that has as many dimensions as the GRU has hidden units. An extensive search was conducted in terms of finding the most effective model parameters for each data subset. Each model was tested with variations on the number of intermediate layers in the predicting feed-forward network (1, 2 or 3), the addition of dropout, whether the GRU was bidirectional, and the number of hidden units in each layer (*large* or *small*, where *large* has approximately 4 times as many hidden units in each layer as *small*). This creates a total of 24 trials per data-subset. All the models were created with Pytorch (Paszke et al., 2019), and each model was trained for 600 epochs using SGD as an optimizer, learning rate = 0.01, momentum = 0.9, L2 regularization with $\lambda = 0.0001$, batch size of 20,

a Cosine Annealing learning rate scheduler, and cross entropy loss. Each layer with the exception of the final layer uses the ReLU activation function. Additionally, training scripts were run using the CPU of a p2.xlarge Elastic Compute Cloud instance provided by Amazon Web Services³.

A summary of the hyperparameters used for each of the best performing models reported in Sec. 5.1 are provided in Tab. 6. The number of hidden units in the GRU is indicated in the column indicating whether the GRU was bidirectional, and the number of hidden units in each layer of the predicting network is indicated next to the column indicating the number of layers.

While selecting the best performing model for DB-C1, DB-C2, DB-C3, and DB-Utt, a model was only considered if it was able to meet or exceed the specificity achieved by a model that was once SOTA, 28.8% (Di Palo and Parde, 2019).

B.2 Transcript Classification

We use a Random Forest (100 trees), Gradient Boosting Estimator (with 150 estimators), SVM (with RBF kernel), a 2-layer neural network (NN, 10 units, Adam optimizer, 200 epochs with learning rate initialized to 0.01), and an ensemble of all 4 aforementioned classifiers (Ens) (Pedregosa et al., 2011). For each extending feature set described in Sec. 3.2, we jointly optimize the number of features selected from the extending feature set (feature selection with k=3, 5, 7, 9, 11, 13, 15, 20,

³<https://aws.amazon.com/ec2/>

25, 30, and all features), whether or not we use the oversampling method SMOTE([Chawla et al., 2002](#)), and the model type used. For feature selection on the *Original* feature set, we attempted feature selection with $k=20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 150, 200, 250, 300,$ and 350. The best performing configuration for each extending feature set is recorded in [Tab. 7](#).