Published in final edited form as:

Speech Commun. 2023 November ; 155: . doi:10.1016/j.specom.2023.102990.

Corresponding author: David Clark , 355 W 16th Street , Suite 4020 , Indianapolis, IN 46202 , Fax: +1 317-963-4916 , clarkdg@iu.edu.

Author contributions

Use this form to specify the contribution of each author of your manuscript. A distinction is made between five types of contributions: Conceived and designed the analysis; Collected the data; Contributed data or analysis tools; Performed the analysis; Wrote the paper. For each author of your manuscript, please indicate the types of contributions the author has made. An author may have made more than one type of contribution. Optionally, for each contribution type, you may specify the contribution of an author in more detail by providing a one-sentence statement in which the contribution is summarized. In the case of an author who contributed to performing the analysis, the author's contribution for instance could be specified in more detail as 'Performed the computer simulations', 'Performed the statistical analysis', or 'Performed the text mining analysis'.

If an author has made a contribution that is not covered by the five pre-defined contribution types, then please choose 'Other contribution' and provide a one-sentence statement summarizing the author's contribution.

Author 1: Justin Bushnell

☒ Conceived and designed the analysis
Specify contribution in more detail (optional; no more than one sentence)
☐ Collected the data
Specify contribution in more detail (optional; no more than one sentence)
☒ Contributed data or analysis tools
Specify contribution in more detail (optional; no more than one sentence)
☒ Performed the analysis
Specify contribution in more detail (optional; no more than one sentence)
☒ Wrote the paper
Specify contribution in more detail (optional; no more than one sentence)
☐ Other contribution
Specify contribution in more detail (required; no more than one sentence)

Author 2: Frederick Unverzagt

☐ Conceived and designed the analysis
Specify contribution in more detail (optional; no more than one sentence)
☒ Collected the data
Specify contribution in more detail (optional; no more than one sentence)
☐ Contributed data or analysis tools
Specify contribution in more detail (optional; no more than one sentence)
☐ Performed the analysis
Specify contribution in more detail (optional; no more than one sentence)
☐ Wrote the paper
Specify contribution in more detail (optional; no more than one sentence)
☐ Other contribution
Specify contribution in more detail (required; no more than one sentence)

Author 3: Virginia G. Wadley

☐ Conceived and designed the analysis
Specify contribution in more detail (optional; no more than one sentence)
☒ Collected the data
Specify contribution in more detail (optional; no more than one sentence)
☐ Contributed data or analysis tools
Specify contribution in more detail (optional; no more than one sentence)
☐ Performed the analysis
Specify contribution in more detail (optional; no more than one sentence)
☐ Wrote the paper
Specify contribution in more detail (optional; no more than one sentence)
☐ Other contribution
Specify contribution in more detail (required; no more than one sentence)

Author 4: Richard Kennedy

☐ Conceived and designed the analysis
Specify contribution in more detail (optional; no more than one sentence)
☒ Collected the data
Specify contribution in more detail (optional; no more than one sentence)
☐ Contributed data or analysis tools
Specify contribution in more detail (optional; no more than one sentence)
☐ Performed the analysis
Specify contribution in more detail (optional; no more than one sentence)
☐ Wrote the paper
Specify contribution in more detail (optional; no more than one sentence)
☐ Other contribution
Specify contribution in more detail (required; no more than one sentence)

Author 5: John Del Gaizo

☐ Conceived and designed the analysis

Specify contribution in more detail (optional; no more than one sentence)
☒ Collected the data
Specify contribution in more detail (optional; no more than one sentence)
☒ Contributed data or analysis tools
Specify contribution in more detail (optional; no more than one sentence)
☐ Performed the analysis
Specify contribution in more detail (optional; no more than one sentence)
☐ Wrote the paper
Specify contribution in more detail (optional; no more than one sentence)
☐ Other contribution
Specify contribution in more detail (required; no more than one sentence)
Author 6: David Glenn Clark
☒ Conceived and designed the analysis
Specify contribution in more detail (optional; no more than one sentence)
☒ Collected the data
Specify contribution in more detail (optional; no more than one sentence)
☒ Contributed data or analysis tools
Specify contribution in more detail (optional; no more than one sentence)
☒ Performed the analysis
Specify contribution in more detail (optional; no more than one sentence)
☒ Wrote the paper
Specify contribution in more detail (optional; no more than one sentence)
☐ Other contribution
Specify contribution in more detail (required; no more than one sentence)
Author 7: Enter author name
☐ Conceived and designed the analysis
Specify contribution in more detail (optional; no more than one sentence)
☐ Collected the data
Specify contribution in more detail (optional; no more than one sentence)
☐ Contributed data or analysis tools
Specify contribution in more detail (optional; no more than one sentence)
☐ Performed the analysis
Specify contribution in more detail (optional; no more than one sentence)
☐ Wrote the paper
Specify contribution in more detail (optional; no more than one sentence)
☐ Other contribution
Specify contribution in more detail (required; no more than one sentence)
Author 8: Enter author name
☐ Conceived and designed the analysis
Specify contribution in more detail (optional; no more than one sentence)
☐ Collected the data
Specify contribution in more detail (optional; no more than one sentence)
☐ Contributed data or analysis tools
Specify contribution in more detail (optional; no more than one sentence)
☐ Performed the analysis
Specify contribution in more detail (optional; no more than one sentence)
☐ Wrote the paper
Specify contribution in more detail (optional; no more than one sentence)
☐ Other contribution
Specify contribution in more detail (required; no more than one sentence)
Author 9: Enter author name
☐ Conceived and designed the analysis
Specify contribution in more detail (optional; no more than one sentence)
☐ Collected the data
Specify contribution in more detail (optional; no more than one sentence)
☐ Contributed data or analysis tools
Specify contribution in more detail (optional; no more than one sentence)
☐ Performed the analysis
Specify contribution in more detail (optional; no more than one sentence)
☐ Wrote the paper
Specify contribution in more detail (optional; no more than one sentence)
☐ Other contribution
Specify contribution in more detail (required; no more than one sentence)
Author 10: Enter author name
☐ Conceived and designed the analysis
Specify contribution in more detail (optional; no more than one sentence)
☐ Collected the data
Specify contribution in more detail (optional; no more than one sentence)

Post-Processing Automatic Transcriptions with Machine

Learning for Verbal Fluency Scoring

Justin Bushnell[1], **Frederick Unverzagt**[2], **Virginia G. Wadley**[3], **Richard Kennedy**[3], **John Del Gaizo**[4], **David Glenn Clark**[1]

[1]Department of Neurology, Indiana University, Indianapolis, IN, USA

[2]Department of Psychology, Indiana University, Indianapolis, IN, USA

[3]Department of Medicine, University of Alabama at Birmingham, Birmingham, AL, USA

[4]Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC, USA

Abstract

Objective: To compare verbal fluency scores derived from manual transcriptions to those obtained using automatic speech recognition enhanced with machine learning classifiers.

Methods: Using Amazon Web Services, we automatically transcribed verbal fluency recordings from 1400 individuals who performed both animal and letter F verbal fluency tasks. We manually adjusted timings and contents of the automatic transcriptions to obtain “gold standard” transcriptions. To make automatic scoring possible, we trained machine learning classifiers to discern between valid and invalid utterances. We then calculated and compared verbal fluency scores from the manual and automatic transcriptions.

Results: For both animal and letter fluency tasks, we achieved good separation of valid versus invalid utterances. Verbal fluency scores calculated based on automatic transcriptions showed high correlation with those calculated after manual correction.

Conclusion: Many techniques for scoring verbal fluency word lists require accurate transcriptions with word timings. We show that machine learning methods can be applied to improve off-the-shelf ASR for this purpose. These automatically derived scores may be satisfactory for some applications. Low correlations among some of the scores indicate the need for improvement in automatic speech recognition before a fully automatic approach can be reliably implemented.

☐ Contributed data or analysis tools
Specify contribution in more detail (optional; no more than one sentence)
☐ Performed the analysis
Specify contribution in more detail (optional; no more than one sentence)
☐ Wrote the paper
Specify contribution in more detail (optional; no more than one sentence)
☐ Other contribution
Specify contribution in more detail (required; no more than one sentence)

Declaration of interests
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Keywords

Automatic speech recognition; cognitive science; verbal fluency; language; machine learning; dementia

1. Introduction

Desiderata for remote screening methods for dementia include efficiency, low cost, wide accessibility, and the possibility of frequent repetition. A sufficiently accurate technique would identify individuals in need of further clinical or neuropsychological evaluations, biofluid testing, or imaging. Incorporating automatic methods such as automatic speech recognition (ASR) into neuropsychological tests can facilitate the processing and analysis of large amounts of data, allowing for the possibility of rapidly and cheaply stratifying individuals on the basis of dementia risk. Some research has shown utility of incorporating ASR to automatically calculate scores from verbal fluency (König et al., 2018; Pakhomov et al., 2015; Tröger et al., 2018), logical memory tests (Lehr et al., 2012), and spontaneous speech (Tóth et al., 2017). Though these studies show that scores derived from ASR perform similarly compared to manually derived scores, complex training or post processing is often needed in order to obtain accurate transcriptions, which may be due to the relatively low sample sizes ($n \leq 165$). Furthermore, there is room for investigating the prognostic value of automatically derived scores as the studies mentioned above only classify patients in terms of severity of cognitive impairment. Finally, previous studies have investigated ASR scores in languages other than English. Replication of such studies in English is important to determine which features are independent of the languages spoken. Here, we evaluate the performance of off-the-shelf automatic speech recognition enhanced with machine learning classifiers to calculate an assortment of verbal fluency scores and compare them to manually derived scores.

The neuropsychological test under study in this work is the verbal fluency task (VFT). VFTs are brief cognitive exams where an individual is instructed to list words that belong to a specific category, usually within a 60 second time frame. The most common types of verbal fluency tasks are semantic (e.g., animals) or initial letter (e.g., letter “F”). The standard metric of VFTs is the total count of valid responses, or raw score. Researchers and clinicians also quantify errors such as intrusions and repetitions. Raw scores and error counts distinguish many varieties of cognitive decline (Butters et al., 1987; Monsch et al., 1992; Obeso et al., 2012; Perez et al., 2020).

Some novel methods of scoring VFTs require measurement of valid word response times. One such method involves the coarse representation of decline in word production over time by segregating raw counts into time intervals. Compared with normal control (NC) and Alzheimer’s disease (AD) groups, individuals with vascular dementia (VAD) or Parkinson disease (PD) exhibit disproportionately greater word production in the first quartile (15s) epoch among all letter fluency cues (A, F, S) (Lamar et al., 2002). Sextile (10s) epochs may also be used for partitioning the task. Raw score of the first three sextile intervals of semantic fluency show additive value in explaining differences among AD and mild

cognitive impairment (MCI), with raw score of the 3rd interval of both semantic and letter fluency also being complementary in predicting MCI and subjective cognitive impairment (SCI) (Fernaeus et al., 2008). Similar results by Linz (2019) show significantly lower total output and lower-frequency words in the third interval when comparing MCI with SCI and healthy controls (Linz et al., 2019).

Other scores have been drawn from the observation that VFT performance relies on separable "clustering" and "switching" components. These components define a pattern similar to foraging in which an individual will repeatedly exploit a sub-category (clustering) and change sub-categories when yield is diminished (switching) (Troyer et al., 1997). Various methods for calculating clustering and switching scores employ the use of semantics, phonology, or orthography (Bushnell et al., 2022; Goñi et al., 2011; König et al., 2018). In this work, we employ the slope difference technique for identifying clusters and switches based on word latencies (i.e., onset times) (Bousfield & Sedgewick, 1944; Gruenewald & Lockhead, 1980). Previous work by our group shows that, among the well-known methods used to calculate clustering and switching scores, timings derived from slope difference with animal fluency are among the best predictors of progressive cognitive impairment (Bushnell et al., 2022). Further, clustering scores as defined by the slope difference algorithm differentiate carriers of apolipoprotein E $\varepsilon 4$ from non-carriers, who produce larger clusters and access those clusters faster (Rosen et al., 2005). Because the slope difference technique requires only latencies, the actual words generated in the VFT do not need to be explicitly recognized, making it well-suited for automatic transcriptions, if valid and invalid utterances may be discerned. While traditional methods of calculating clustering scores require recognition of links between non-sequential words, the slope-difference approach can only identify linkages between consecutive words. We refer to sequences of linked words as chains, as the term cluster in related work is somewhat more restrictive.

The elapsed times between sequential valid words, known as inter-word intervals (IWIs), have been another source from which to draw novel scores. In previous work, we used IWIs to derive a speed score; this is a modified raw score in which a higher number represents more words with faster response times. In survival analyses evaluating time to incident cognitive impairment, speed scores make independent contributions to variance and improve concordance (Ayers et al., 2022). Other work employs a simple regression model to predict IWIs from word indices (numerical order in which words are listed). According to Mayr (2002) and Mayr and Kliegl (2000), the slope coefficient of the regression measures the semantic search that increases in difficulty as time elapses, resulting in longer IWI with each produced word. The intercept of the model indexes time-invariant features of the task that relate to its difficulty (both task difficulty and factors specific to the individual, such as formulation and adherence to a search strategy and psychomotor speed). Utilizing this same technique to calculate regression parameters, we found in previous work that the intercept has potential value for predicting acute decline in cognition (Bushnell et al., 2022).

Here, we investigate the feasibility of calculating verbal fluency scores directly from the telephone-administered raw audio in both animal and letter fluency tasks. We obtained automatic transcriptions from Amazon Web Services (AWS). We manually corrected the

timings and contents of each utterance to generate “gold standard” transcriptions for comparison. Through the use of machine learning, we identified valid utterances in the AWS transcriptions. We automatically calculated verbal fluency scores on both sets of transcriptions. For the manual transcriptions, we used utterances manually selected as valid. For the automatic transcriptions, we used utterances designated as valid by the machine learning classifiers. Using correlations, we compared the scores from the two sets of transcriptions.

2. Methods

See Figure 1 for a diagram showing the steps in the analysis.

2.1 Subjects

Utilizing the same dataset employed in two other published analyses (Ayers et al., 2022; Bushnell et al., 2022), we selected 703 cases of incident cognitive impairment (ICI) from the Reasons for Geographic and Racial Differences in Stroke (REGARDS) study (Howard, 2013; Howard et al., 2005; Wadley et al., 2011). We determined ICI on the basis of longitudinal scores of the Six-Item Screener (SIS), a brief cognitive exam with three temporal orientation questions and a three-word delayed recall (Callahan et al., 2002). Each case was matched to a control according to age, sex, race, geographic region, and education for a total of 1406 subjects. All subjects performed both animal and letter F fluency tasks at baseline prior to detection of any cognitive impairment. Participants were stratified by fluency task performance to facilitate partitioning them into comparable training and test sets. The size of the training and test sets were 1200 and 206, respectively. Because some scores can only be calculated based on multiple responses, we then removed any subject whose actual raw score was below 2 for either verbal fluency task. The training and test sets were reduced to 1196 and 204, respectively. Within the training set, a validation set with 200 subjects was used for pretraining.

2.2 Data pre-processing

2.2.1 Transcriptions—We obtained a preliminary transcription of each verbal fluency recording through AWS. We provided AWS with lists of candidate animals and F-words to facilitate accurate transcriptions (both lists are available by request from the corresponding author). To obtain manual transcriptions, we used the ASR transcriptions as a basis and implemented a series of steps to create accurate transcriptions. First, we employed a voice activity detector (VAD) to identify areas in which the ASR failed to detect speech for review. Next, we utilized a custom MATLAB (Natick, MA) transcription tool to manually correct both the timings and contents of the ASR transcriptions, also reviewing the missed utterances identified by the VAD. Our transcription tool displays the audio waveform and permits selection of a window for audio playback. A transcription for a segment of audio may then be entered and is displayed above the waveform. Utterance timings are automatically extracted based on the selection window. A human rater listened to every audible portion of each recording. Utterances that were difficult to understand were resolved through informal consensus among the study personnel. The first author (JB) performed extensive quality control, making changes or consulting the lead author

(DGC) for questionable items. Test administrator utterances were identified and marked. For automatic identification of repetitions, we created ancillary lists in which each valid word was paired with a canonical form (e.g., “hippos” was paired with “hippopotamus”). Within a single list, we considered any second instance of a valid word with the same canonical form to be a repetition. Intrusions were identified on a per-transcription basis and marked as such in the manual transcriptions. For the automatic transcriptions, we presumed that intrusions were filtered out by the classifiers (see section 2.3), but we could not automatically count them.

2.2.2 Alignment—To label utterances from the original ASR transcriptions as valid or invalid, we first aligned them with the manually transcribed words on the basis of timings. Iterating through each known valid word within the manually corrected transcriptions, we considered an utterance from the ASR transcription to be aligned with an utterance from the corrected transcription if at least 50% of the briefer utterance overlapped the time interval of the longer utterance. If multiple ASR words were aligned with the same valid word, the ASR words were then concatenated and considered as a single utterance. This step allowed us to place nonsensical sequences from the ASR transcription, such as “Iraq tuna,” into alignment with valid material (“a raccoon”). ASR material that aligned with valid utterances was labeled as valid while unaligned material was labeled as invalid. We then concatenated any sequential, invalid ASR words into a single utterance if they were contiguous in time (i.e., the end time of the first was identical to the start time of the second). For sample alignments, see tables e1 – e4 in the supplementary material.

2.3 Identification of Valid Utterances and Repetitions

We trained Naïve Bayes (NB), random forest (RF), and support vector machine (SVM) classifiers to discern between valid and invalid utterances. To train and apply these classifiers, we generated four different feature matrices. Thus, for each fluency task, we fit 12 classifiers (3 architectures × 4 feature matrices). Classifiers were fit using R, version 3.5.1 (R-Core-Team, 2018).

The first feature matrix consisted of a simple type-token matrix. The rows of the matrix represented the ASR utterances of all transcriptions, while the columns represented the set of all transcribed words from the ASR (types). A numerical entry in the matrix indicated the number of times a given word occurred in an utterance (tokens). The second feature matrix was a term frequency-inverse document frequency (tf-idf) matrix derived through a simple transformation of the first matrix (Jurafasky & Martin, 2009). The third feature matrix consisted of phonological overlap measurements between each utterance and each feature. Because utterances often consisted of more than one transcribed word, we evaluated the phonological overlap between any single word within an utterance and each possible feature. In addition to single words, we also examined the overlap between the entire utterance and each feature. The matrix entry for a given row (utterance) and column (feature) was the maximum of the phonological overlap scores thus obtained. The fourth feature matrix consisted of a simple element-wise sum of the tf-idf and phonological overlap matrices.

We fit the NB classifiers using the naivebayes library (0.9.7) for R (Majka, 2019). The hyperparameters included Laplace smoothing of 1.0 and use of a Poisson distribution for predictor variables consisting only of positive integers. For the RF classifier, we employed the ranger library (0.13.1) (Wright & Ziegler, 2017). Each RF classifier included 500 trees, placing no limit on depth, with *mtry* (number of features to randomly select for consideration at each branch point) set equal to the square root of the number of features. We fit the SVM classifiers using the e1071 library (1.7.9) (Meyer & Wien, 2015), setting both the cost and gamma hyperparameters to 1.0.

For animal fluency, the training set matrices (with the validation set included) contained 33,172 rows. The test set matrices contained 5,732 rows. Both matrices contained 2,417 columns. For letter fluency, the training set matrices (with the validation set included) contained 28,451 rows. The test set matrices contained 4,762. Both matrices contained 2,533 columns. It should be noted that for training the classifier we excluded any manually transcribed word (i.e., valid VFT word) in which there was not an aligned ASR utterance. (Because we were training based on the ASR transcription, there was no way to include information that was missing). We report ASR quality scores in the results section. We removed material marked as invalid by the classifier prior to calculating verbal fluency scores from the ASR transcriptions. Using the same ancillary lists for identifying repetitions in the manual transcriptions (see section 2.2.1), we marked and removed repetitions if a previous word matched any form (e.g., inflection) of the current word. Upon removing repetitions, we obtained the final version of the automated transcriptions. We plotted receiver operating characteristic (ROC) curves and calculated area under the ROC curve (AUC).

For each VFT and each feature matrix, we generated two ensemble classifiers by combining the output from the NB, RF, and SVM classifiers. The first ensemble consisted of a simple average of output from the three classifier architectures. The second ensemble consisted of a *weighted* average, in which the weights were derived by fitting a simple logistic regression with the output from the three classifiers as independent variables and validity as the dependent variable.

2.4 Evaluating and Improving ASR Accuracies

To assess the possibility of calculating scores based on word semantics (e.g., traditional clustering and switching scores), we calculated three types of word error rates (WER) using the formula WER = $(S+D+I)/N$, where S is substitutions, D is deletions, I is insertions, and N is the number of tokens based on the manual transcriptions. First, we report the standard WER derived from our aligned automatic and manual transcriptions. Second, Following König et al. (2018), we calculated the verbal fluency error rate (VFER) which quantifies word-level errors based only on valid verbal fluency words. Initially, we converted each manually transcribed word to the canonical form and ASR utterance where possible. To simplify the calculation of number of errors (numerator), we then calculated the difference of the word list sets, using each set as the reference to obtain two sets for each row in the aligned transcription. The set with the maximum length represents the total number of errors between the ASR aligned utterance and the manually transcribed utterance. If the

ASR software missed the word altogether, the VFER for the aligned utterance was weighted by the total number of words missed. The denominator is the total number of words in the manually transcribed utterance.

We calculated a third type of WER which we call phonemic error rate (PhER). This error rate captures phonemic-level similarities between the manual and automatic transcriptions, providing additional information for how proximate the ASR software was in transcribing. Here, we left both the ASR and manual utterances in their original, potentially inflective state. Initially, we phonetically aligned the manually transcribed utterance with their aligned ASR utterance. We then counted the number of instances where the phonemes were not identical (substitutions) or where one phoneme sequence had a phoneme not present in the other (insertions and deletions). Similar to the VFER, the PhER for the aligned utterance was weighted by the total number of phonemes missed if the ASR software missed the word altogether. We then divided the number of phoneme errors by the total number of phonemes in the manually corrected transcription. For details of the phonological alignment algorithm, see supplementary materials of Bushnell et al (2022).

In order to improve the VFER, we implemented a phonologically based algorithm (identical to the alignment for PhER) to map each aligned ASR utterance to its closest phonological neighbor from the list of valid verbal fluency words. Next, we converted the mapped words and the human transcribed words to their canonical form. Finally, we recalculated VFER.

2.5 Verbal Fluency Scores

2.5.1 Raw Scores—Raw score was calculated by counting the number of utterances in each automatic transcription (after processing). Along with total raw score, we calculated raw scores within six 10s (sextile) epochs. The epoch to which a word belonged was based on onset time.

2.5.2 Clustering and Switching Scores—We computed clustering and switching scores by means of the slope difference algorithm, a time-based method that relies only on word onset times and indices. Following Bousfield and Sedgewick (1944), we plotted each subject's verbal fluency task using the onset times as the independent variable and the index of the onset times as the dependent variable. To this reference curve, we fit an exponential curve to the latencies of valid words following the formula $y = c(1 - e^{-mt})$, where c and

$m$are parameters and t represents the onset times of the valid words. Using a custom MATLAB program, we determined values of c and m to minimize the sum of squared deviations between the exponential curve and the reference curve. To determine if sequential words are related, the program compares the slope between the exponential curve and reference curve at the halfway point between the two words. If a word is produced more quickly than predicted by the fitted exponential curve, the slope difference is positive, and the words are considered "linked." A negative slope difference indicates absence of a link, or a "switch." Once the links are established, we count the number of switches and measure the lengths of the chains. We report the average of these lengths as a type of clustering score.

2.5.3 Inter-word Intervals—Following Mayr (2002) and Mayr and Kliegl (2000), we fit a simple linear regression to every VFT, using IWI as the dependent variable and the index of the IWI as the independent variable ($IWI \sim sx + b$). Only IWIs that occurred between consecutive, valid words were considered.

Another score computed from the set of IWIs is the speed score. These scores were computed by taking the 4^{th} root of all IWIs, placing them on the interval [0,1] with min-max normalization, and subtracting the result from 1.0 (such that faster transitions receive higher scores). An individual's speed score was the sum of these transformed IWIs. Following Bushnell et al. (2022), we also calculated speed scores separately for switch and edge transitions as determined by the slope difference algorithm.

2.6 Statistical Analyses

We compared scores derived from the processed ASR transcriptions (i.e., utterances determined to be valid by the best classifier and their onset times) to those derived from the manually corrected transcriptions. The main comparison was Pearson correlation. Means and standard deviations were calculated on the signed and unsigned (i.e., absolute value) differences between the estimated and actual scores (Actual – Estimated). Because the c and m parameter ranges were extreme, we used Spearman correlation and summarized central tendency and dispersion with median and interquartile range. The signed and unsigned differences were performed after applying a z-transformation to each distribution. All statistical analysis were conducted in R.

To rigorously test the utility of our automatically derived scores for clinically relevant predictions, we performed an analysis like the one we presented in Bushnell et al. (2022), using the statisticalrethinking library (McElreath, 2020). We selected the variable sets that yielded the best fit in the previously published work and fit two Bayesian logistic regression models for each variable set: one in which the variables were manually derived and one in which the variables were automatically derived. The "badness of fit" for each model was quantified using the Watanabe-Akaike information criterion (WAIC), and the WAIC values were used to compare each pair of models in terms of "weight." Models that fit the data better receive a higher weight. In addition, we calculated AUC, sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV), and F1 score. We performed the Bayesian logistic regressions on animal and letter fluency tasks separately, then again with the fluency tasks combined.

3. Results

WER for animal fluency was 42.9%, with an average of 2.2 insertions, 6.0 deletions, and 12.0 substitutions per transcription. For letter fluency, WER was 41.8%, with an average of 2.8 insertions, 5.3 deletions, and 11.7 substitutions per transcription. The VFERs for animal and letter fluency were 39.3% and 37.9%, respectively. With the implementation of the phonological mapping algorithm for error correction, the recalculated VFERs were 25.6% and 32.2%. The PhER for animal fluency was 23.7% and for letter fluency was 19.5%. Filled pauses (e.g., "um") were common, consisting of 12.8% of the overall sample

from AWS animal transcriptions and 11.4% of the overall sample from AWS letter F transcriptions.

The RF classifier trained with the combined phonological overlap + tf-idf matrix gave the best AUC for both animal (0.9649) and letter F fluency (0.9579). The balanced accuracy for the identification of repetitions for animal fluency was 0.81 (Pearon's r also 0.81) and was 0.82 (Pearson's r 0.77) for letter fluency.

Pearson's correlations for animal fluency raw score was 0.91 and letter fluency was 0.88. The maximum epoch correlation for animal fluency was 0.90 on the second sextile. The minimum epoch correlation for animal fluency was 0.81 on the fifth sextile. Maximum epoch correlations for letter fluency occurred for the first sextile at 0.87. Minimum epoch correlation for letter fluency was 0.79 on the fourth sextile.

When comparing Bayesian logistic models fit with manual vs. automatically derived scores, the proportions of weight assigned to the automatic animal fluency, letter fluency, and combined Bayesian logistic regression models were 4.5%, 0.9%, and 6.9%, respectively.

4. Discussion

Using automatic methods such as automatic speech recognition for remote-capable neuropsychological tasks could help in the early detection of cognitive decline. Such methods may also aid in directing individuals to the next step in evaluation, an important issue given the potentially broad selection of other tests, all with varying levels of cost and invasiveness. Ongoing improvements in ASR will further improve estimates of scores and may soon converge with those based on manual transcriptions. However, obtaining perfect transcriptions may require additional machine learning techniques for speaker diarization, classification of utterances, and quantification of repetitions and intrusions.

In this study, we enhanced off-the-shelf ASR with machine learning algorithms to identify valid material. We aimed to assess the accuracy of verbal fluency scores between automatically and manually transcribed audio recordings. We processed remotely acquired audio recordings from the REGARDS dataset with ASR to automatically obtain transcriptions for both animal and letter F fluency. To label the ASR utterances as valid or invalid, we first aligned the ASR utterances with valid words in the corrected transcriptions based on time and then trained classifiers to separate valid and invalid utterances. Repetitions were identified, quantified, and removed (for scoring purposes) through comparison of canonical forms of items in each transcribed list. Manually transcribed versions were obtained by correcting both the time and contents of the ASR transcriptions. We then calculated scores for the raw ASR and manually corrected transcriptions. Scores derived from the two types of transcriptions on the test set were then compared.

4.1 ASR Accuracies

VFER rates were somewhat better than WER, suggesting that fewer errors occurred in the valid VFT utterances. VFER was diminished further by our phonological mapping technique but remained sufficiently high (>25%) to suggest that off-the-shelf ASR technology paired

with our method of automatically correcting utterances would be insufficient for calculating scores that require explicit, accurate word identification. The discrepancy between our animal fluency VFER and the error rate acquired by König et al. (2018) is likely due to different ASR technology used in transcribing audio. The ASR software we used did not provide a list of possible candidate words with confidence scores from which to choose, producing a less refined output from the ASR compared with König et al. The extent to which the automatic alignment of utterances contributes to the discrepancy is unknown but may also be a factor in higher VFER.

With our simple phonological mapping algorithm, we observed a marked decrease in VFER for animal fluency and a notable, but smaller, decrease in VFER for letter fluency. The error rate with phonological mapping for semantic verbal fluency suggests that calculating semantic-based scores could be possible as this error rate is close to the error rate achieved by König et al. The VFER for letter fluency suggests that the accuracies may still be insufficient but further investigation is necessary.

The PhER for both fluency tasks indicates that ASR provides a similar sounding word or phrase in place of the correct word. This similarity makes possible the improvements in error rates through phonological mapping.

4.2 *Post hoc* Sensitivity Analysis of Repetitions

The balanced accuracies for the classification of repetitions indicate there is room for improvement. The method we employed is straightforward but depends on the accuracy of the ASR. To see the influence repetitions had on scores, we simulated perfect repetition detection on the valid utterances identified by the best classifier by recalculating all scores using the manually identified repetitions. The average change in correlation across all scores for animal and letter fluency was an increase of 0.01 and 0.025, respectively. The maximum increase in correlation for animal fluency was 0.038 and for letter fluency was 0.081 (both for switch speed). Any improvement in ASR will directly aid the identification of repetitions in a model such as ours. Alternatively, ASR could be further improved with methods from digital signal processing for automatic identification of repetitions. However, for the most part, improving recognition of repetitions will lead only to modest improvements in automatic scoring.

4.3 Classifiers

For both fluency tasks, the best AUC was obtained by fitting an RF classifier with the combined phonological overlap + tf-idf matrix (animal train/test: 0.9809/0.9649; letter F train/test: 0.9732/0.9579). With the exception of using the type-token feature matrix, the RF models outperformed all other classifiers and ensembles.

Though there is nothing inherently different about the identification of letter F words and animal words from an ASR standpoint (as supported by the ASR quality scores in the results section), the slight difference in AUC may be explained by the generally smaller letter F dataset (due to lower raw scores on average) and the larger set of candidate F-words (867 letter F *vs.* 433 animal words).

4.4 Bayesian Logistic Regression Models

The Bayesian logistic regression models show that the ASR-derived scores have similar AUCs compared to the same scores derived from manual transcriptions. However, the weights of each model indicate that the predictive value of the ASR-derived scores is severely weakened when compared to their manual counterparts. We would not expect the best possible automatically derived scores to perform better than manually derived scores. Given perfect automatically derived scores, the models should receive equal WAIC scores, and each model would therefore receive 50% of the weight in the model comparison.

4.5 Interpretation of Results

We find that off-the-shelf ASR, when accompanied by our RF valid-item recognizer and our automatic method for identifying repetitions, yields scores that correlate strongly with those obtained through labor intensive manual transcription. Among the highest correlations for both fluency tasks are raw scores (with epochs), repetitions, and the speed score. Perhaps the simplicity of the calculation required for these scores contributes to the high correlation. The number of switches (according to the slope difference algorithm) and edge speed also showed high correlations for both VFTs. Mean chain lengths of both verbal fluency tasks have low correlations. These scores require correct identification of multiple consecutive words and misidentification of any valid item could split a single chain into separate, smaller chains. Similarly, misidentification of an invalid item could lead to the joining of two short chains. This score would require still higher accuracies of valid word identification to make it useful for detecting or discerning among diseases. Animal M&K (Mayr and Kliegl) correlation scores were low in comparison to other scores, including letter F M&K scores. This inconsistency may be due to outliers within the animal fluency task.

König, et al. (2018) performed a similar comparison of animal fluency scores derived from automatic transcriptions to those from manually corrected transcriptions, with some comparable results. We report very similar correlations for raw score (0.91 compared to 0.92 in König et al.) and number of switches (0.79 vs. 0.80). We did not observe such a strong correlation for the clustering scores we reported (0.57 vs. 0.86), but we employed a different method that only considered the speed with which consecutive words were generated. (In previous work, we observed that this method yielded better predictions of future of cognitive impairment than other clustering methods, but we did not evaluate the method described by König et al.) There were several other key differences between the two studies. First, our participants spoke English rather than French. Thus, to the extent that the results agree, we have replicated König et al.'s work in an additional language. Such replication will be necessary to determine which aspects of this line of research are language-independent and to develop language-specific tools for evaluating cognitive impairment in speakers of different languages. Second, we report similar results for letter fluency as well as animal fluency. Third, apart from having a larger sample size, our participants were recruited in the context of an epidemiological study that sampled individuals of two races across the 48 contiguous United States. The ethnic and geographic heterogeneity might have provided a more conservative evaluation of the capability of current off-the-shelf ASR techniques for this purpose. Fourth, we employed AWS rather than Google Cloud Platform ASR. Although we did acquire transcriptions from Google at a cost similar to that for AWS, we found that

the timings were not sufficiently accurate to use for aligning with our manual transcriptions. Although it might have been possible to perform a purely phonological alignment, we would have been forced to rely on the timings from our manual transcriptions. This approach would not have suited our purpose, as the non-traditional scores we were evaluating depended on accurate timings and we wished to evaluate the ASR system's capacity to measure them. Finally, our data set did not include any individuals who were thought to be impaired at the time the recordings were made. One expects that if two scores are influenced by severity of cognitive impairment, inclusion of cognitively impaired individuals may strengthen the correlation between the two scores. Some of the scatter plots shown in Figure 1 (König et al., 2018). reveal a strong tendency for individuals within a diagnostic category to cluster. This tendency could have strengthened the correlations in that work.

4.6 Study Strengths

There are many robust characteristics of this study. The generous sample size offers the ability to train the classifiers on a large portion of the data and to preliminarily apply the model to a validation set, yielding a more generalizable model with excellent performance on the held-out test set. For each recording, we obtained highly accurate manual transcriptions using software which allowed us to quickly replay less intelligible words, identify test administrator utterances, and mark timings with the help of visualization of each waveform. Our study applied two types of verbal fluency tasks. Similar results were obtained on both VFTs, suggesting that our approach is useful for both tasks. We also showed that a variety of features from the literature can be calculated using this automatic method because precise identification of lexical content is not required for scoring. Finally, the audio used was recorded over the telephone and our results directly pertain to remote administration.

4.7 Limitations

The main limitation of this study is the investigation of only one ASR system. There are numerous speech-to-text packages, likely with varying accuracies in both timings and content. Other packages such as Google Cloud Platform, Microsoft Azure, Dragon, and IBM Watson may perform better or worse than AWS. It may be possible to improve on our results by integrating output from multiple classifiers. Another limitation is that ASR accuracy of our application may be affected by factors of brain health reflected by ICI group membership. For example, individuals with AD have a greater proportion of voice breaks (König et al., 2015; Meilán et al., 2014). Furthermore, with reference to healthy controls, individuals with mild cognitive impairment (MCI) have weaker speech when controlling for age, greater shimmer (instability in amplitude of voice), and greater dysphonia (Themistocleous et al., 2020). The presence of these differences could lead to more errors in ASR transcriptions, yet if they were accurately measured, they could provide additional prognostic features. Because the SIS is simple and brief by design, the ICI group is almost certainly heterogeneous, and the extent of speech irregularities is unknown.

5. Conclusion

Further advances in ASR are needed to perfect the automatic scoring of verbal fluency and other cognitive tests. The value of current off-the-shelf ASR is enhanced by simple machine learning techniques. Much work remains to be done to determine the utility of automatic and timing-based scores for diagnosis and prognosis. The incorporation of multiple ASR packages into an ensemble may yield improvements. We anticipate that analysis of voice quality and response times will add to standard neuropsychological test scores and that ongoing improvements in ASR will lead to widespread adoption of these additional measures.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Funding sources: NIH U01 NS041588; NIH P30 AG010133; P30 AG072976

Funding

The authors have no conflicts of interest or financial disclosures. This work was supported by National Institutes of Health grants to the REGARDS Study (NIH U01 NS041588) and to the Indiana University Alzheimer Disease Research Center (NIH P30 AG010133 and P30 AG072976).

References

Ayers M, Bushnell J, Gao S, Unverzagt F, Del Gaizo J, Wadley V, Kennedy R, & Clark D. (2022). Verbal fluency response times predict incident cognitive impairment. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 14. 10.1002/dad2.12277

Bousfield WA, & Sedgewick CHW (1944). An Analysis of Sequences of Restricted Associative Responses. The Journal of General Psychology, 30(2), 149–165. 10.1080/00221309.1944.10544467

Bushnell J, Svaldi D, Ayers M, Gao S, Unverzagt F, Gaizo J, Wadley V, Kennedy R, Goñi J, & Clark D. (2022). A comparison of techniques for deriving clustering and switching scores from verbal fluency word lists. Frontiers in Psychology, 13. 10.3389/fpsyg.2022.743557

Butters N, Granholm E, Salmon DP, Grant I, & Wolfe J. (1987). Episodic and semantic memory: a comparison of amnesic and demented patients. J Clin Exp Neuropsychol, 9(5), 479–497. 10.1080/01688638708410764 [PubMed: 2959682]

Callahan CM, Unverzagt FW, Hui SL, Perkins AJ, & Hendrie HC (2002). Six-item screener to identify cognitive impairment among potential subjects for clinical research. Med Care, 40(9), 771–781. 10.1097/01.MLR.0000024610.33213.C8 [PubMed: 12218768]

Fernaeus SE, Ostberg P, Hellstrom A, & Wahlund LO (2008). Cut the coda: early fluency intervals predict diagnoses. Cortex, 44(2), 161–169. 10.1016/j.cortex.2006.04.002 [PubMed: 18387545]

Goñi J, Arrondo G, Sepulcre J, Martincorena I, Velez de Mendizabal N, Corominas-Murtra B, Bejarano B, Ardanza-Trevijano S, Peraita H, Wall DP, & Villoslada P. (2011). The semantic organization of the animal category: evidence from semantic verbal fluency and network theory. Cogn Process, 12(2), 183–196. 10.1007/s10339-010-0372-x [PubMed: 20938799]

Gruenewald P, & Lockhead G. (1980). The free recall of category examples. Journal of Experimental Psychology: Human Learning and Memory, 6, 225–240. 10.1037/0278-7393.6.3.225

Howard V. (2013). Reasons Underlying Racial Differences in Stroke Incidence and Mortality. Stroke, 44, S126–S128. 10.1161/STROKEAHA.111.000691 [PubMed: 23709708]

Howard VJ, Cushman M, Pulley L, Gomez CR, Go RC, Prineas RJ, Graham A, Moy CS, & Howard G. (2005). The reasons for geographic and racial differences in stroke study: objectives and design. Neuroepidemiology, 25(3), 135–143. 10.1159/000086678 [PubMed: 15990444]

Jurafasky D, & Martin JH (2009). Speech and Language Processing (Second ed.). PrenticeHall, Inc.

König A, Linz N, Tröger J, Wolters M, Alexandersson J, & Robert P. (2018). Fully Automatic Speech-Based Analysis of the Semantic Verbal Fluency Task. Dementia and Geriatric Cognitive Disorders, 45, 198–209. 10.1159/000487852 [PubMed: 29886493]

König A, Satt A, Sorin A, Hoory R, Toledo-Ronen O, Derreumaux A, Manera V, Verhey F, Aalten P, Robert P, & David R. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 1, 112–124. 10.1016/j.dadm.2014.11.012

Lamar M, Price CC, Davis KL, Kaplan E, & Libon DJ (2002). Capacity to maintain mental set in dementia. Neuropsychologia, 40(4), 435–445. 10.1016/s0028-3932(01)00125-7 [PubMed: 11684176]

Lehr M, Prud'hommeaux E, Shafran I, & Roark B. (2012). Fully Automated Neuropsychological Assessment for Detecting Mild Cognitive Impairment. Proceedings of the 13th Annual Conference of the International Speech Communication Association., 2.

Linz N, Lundholm Fors K, Lindsay H, Eckerström M, Alexandersson J, & Kokkinakis D. (2019). Temporal Analysis of the Semantic Verbal Fluency Task in Persons with Subjective and Mild Cognitive Impairment. 10.18653/v1/W19-3012

Majka M. (2019). naivebayes: High Performance Implementation of the Naïve Bayes Algorithm. In https://CRAN.R-project.org/package=naivebayes

Mayr U. (2002). On the dissociation between clustering and switching in verbal fluency: Comment on Troyer, Moscovitch, Winocur, Alexander and Stuss. Neuropsychologia, 40, 562–566. 10.1016/S0028-3932(01)00132-4 [PubMed: 11749985]

Mayr U, & Kliegl R. (2000). Complex semantic processing in old age: does it stay or does it go? Psychol Aging, 15(1), 29–43. [PubMed: 10755287]

McElreath R. (2020). Statistical Rethinking. A Bayesian course with examples in R and Stan. Taylor & Francis.

Meilán J, Martínez-Sánchez F, Carro J, López D, Millian Morell L, & Arana J. (2014). Speech in Alzheimer's Disease: Can Temporal and Acoustic Parameters Discriminate Dementia? Dementia and geriatric cognitive disorders, 37, 327–334. 10.1159/000356726 [PubMed: 24481220]

Meyer D, & Wien FT (2015). Support Vector Machines. The Interface to Libsvm in Package e1071.

Monsch AU, Bondi MW, Butters N, Salmon DP, Katzman R, & Thal LJ (1992). Comparisons of verbal fluency tasks in the detection of dementia of the Alzheimer type. Arch Neurol, 49(12), 1253–1258. 10.1001/archneur.1992.00530360051017 [PubMed: 1449404]

Obeso I, Casabona E, Bringas ML, Alvarez L, & Jahanshahi M. (2012). Semantic and phonemic verbal fluency in Parkinson's disease: Influence of clinical and demographic variables. Behav Neurol, 25(2), 111–118. https://www.ncbi.nlm.nih.gov/pubmed/22530265 [PubMed: 22530265]

Pakhomov S, Marino S, Banks S, & Bernick C. (2015). Using Automatic Speech Recognition to Assess Spoken Responses to Cognitive Tests of Semantic Verbal Fluency. Speech Communication, 75. 10.1016/j.specom.2015.09.010

Perez M, Amayra I, Lazaro E, Garcia M, Martinez O, Caballero P, Berrocoso S, Lopez-Paz JF, Al-Rashaida M, Rodriguez AA, Luna P, & Varona L. (2020). Intrusion errors during verbal fluency task in amyotrophic lateral sclerosis. PLoS One, 15(5), e0233349. 10.1371/journal.pone.0233349

R-Core-Team. (2018). R: a language and environment for statistical computing. In http://www.R-project.org/

Rosen V, Sunderland T, Levy J, Harwell A, McGee L, Hammond C, Bhupali D, Putnam K, Bergeson J, & Lefkowitz C. (2005). Apolipoprotein E and category fluency: Evidence for reduced semantic access in healthy normal controls at risk for developing Alzheimer's disease. Neuropsychologia, 43, 647–658. 10.1016/j.neuropsychologia.2004.06.022 [PubMed: 15716154]

Themistocleous C, Eckerstrom M, & Kokkinakis D. (2020). Voice quality and speech fluency distinguish individuals with Mild Cognitive Impairment from Healthy Controls. PLoS One, 15(7), e0236009. 10.1371/journal.pone.0236009

Tóth L, Hoffmann I, Gosztolya G, Vincze V, Szatloczki G, Bánréti Z, Pakaski M, & Kalman J. (2017). A Speech Recognition-based Solution for the Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech. Current Alzheimer Research, 14. 10.2174/1567205014666171121114930

Tröger J, Linz N, König A, Robert P, & Alexandersson J. (2018). Telephone-based Dementia Screening I: Automated Semantic Verbal Fluency Assessment. 10.1145/3240925.3240943

Troyer AK, Moscovitch M, & Winocur G. (1997). Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. Neuropsychology, 11(1), 138–146. 10.1037//0894-4105.11.1.138 [PubMed: 9055277]

Wadley V, Unverzagt F, McGuire L, Moy C, Go R, Kissela B, McClure L, Crowe M, Howard V, & Howard G. (2011). Incident cognitive impairment is elevated in the stroke belt: The REGARDS Study. Annals of neurology, 70, 229–236. 10.1002/ana.22432 [PubMed: 21618586]

Wright MN, & Ziegler A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, 77(1). 10.18637/jss.v077.i01

Highlights

- We enhanced off-the-shelf ASR with classifiers to score verbal fluency tasks
- Many novel scores utilizing timings of words can be calculated automatically
- We achieved high AUCs for the identification of valid words
- Most automated scores correlate strongly with manual scores

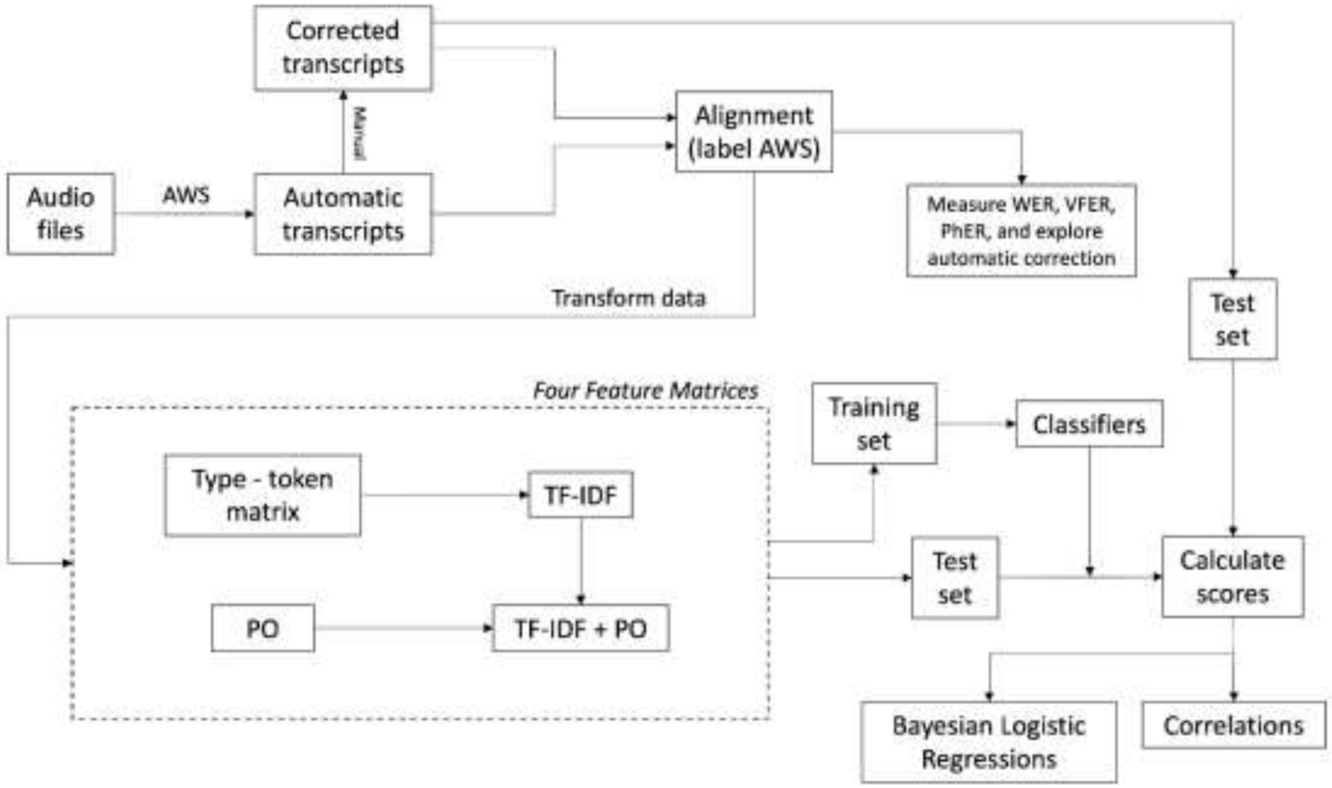

Figure 1: Logical flow of analysis

AWS=Amazon Web Services Transcribe tool, PhER=phonological error rate, PO=phonological overlap, TF-IDF=term frequency-inverse, VFER=verbal fluency error rate, WER=word error rate.

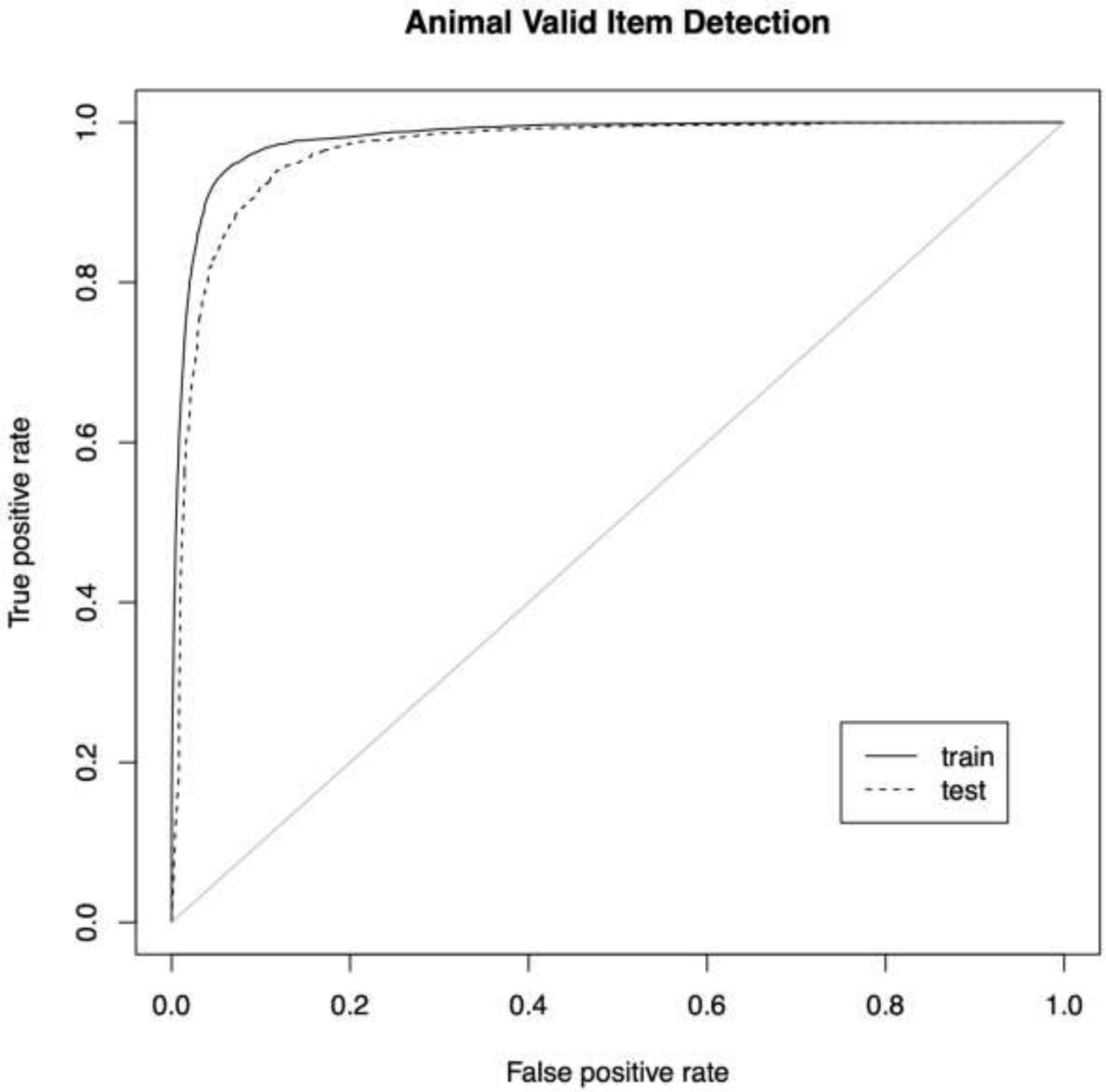

Figure 2: ROC Animal

The receiver operating characteristic (ROC) curve for the detection of valid utterances by the best (random forests) classifier.

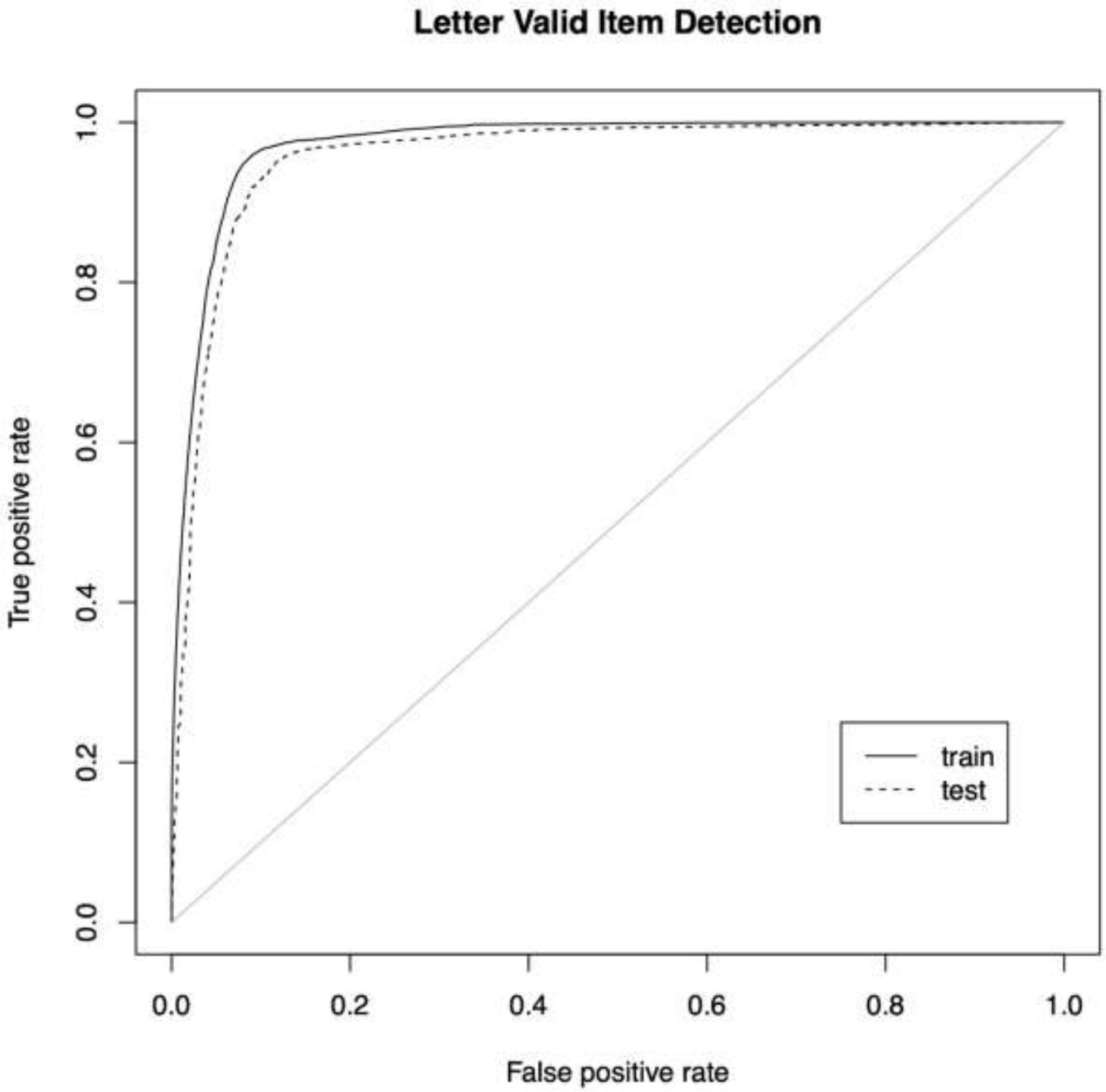

Figure 3: ROC Letter

The receiver operating characteristic (ROC) curve for the detection of valid utterances by the best (random forests) classifier.

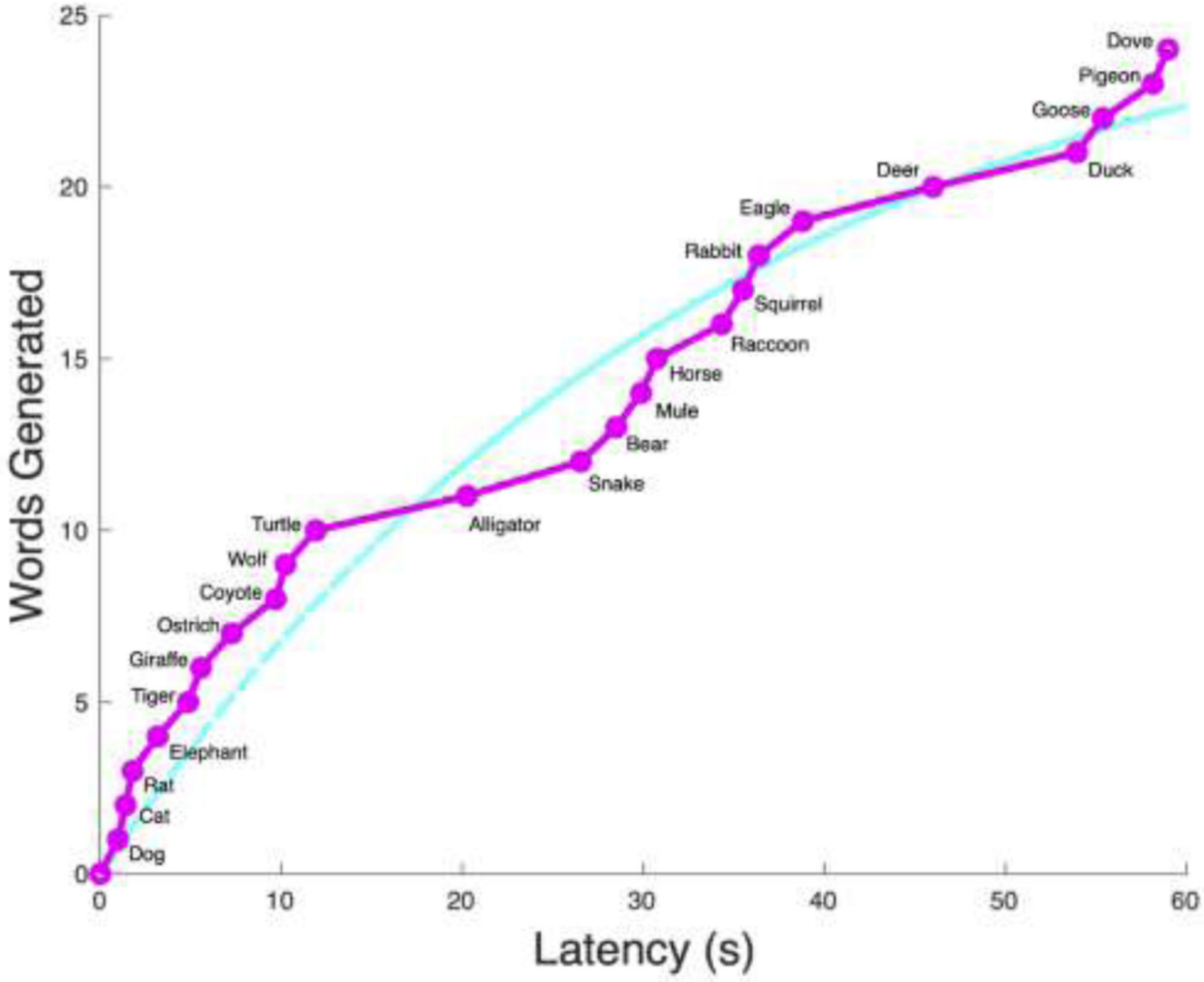

Figure 4: Slope Difference Animal

The jagged magenta curve represents the total number of words generated over the 60 second time period. The smooth cyan curve is the fitted exponential curve. The chains of animals linked based on positive slope difference are: [dog, cat, rat, elephant], [tiger, giraffe], [ostrich], [coyote, wolf, turtle], [alligator], [snake, bear, mule, horse], [raccoon, squirrel, rabbit, eagle], [deer], [duck, goose, pigeon, dove]. The switching score is 8 and the mean chain size is 2.67.

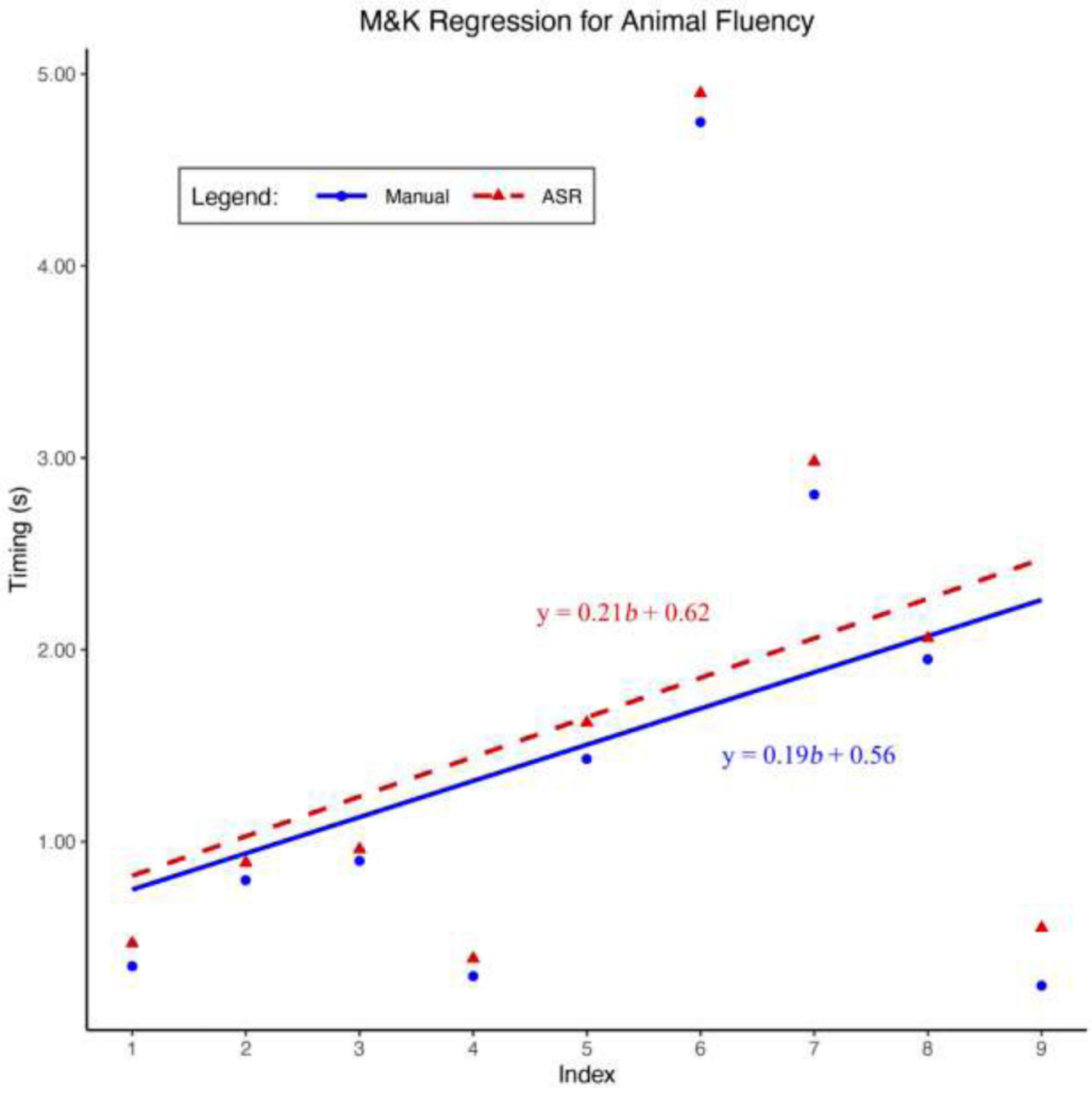

Figure 5: Mayr and Kliegl Animal Regression

Shown are the Mayr and Kliegl linear regressions using the estimated transcriptions and manual transcriptions from a sample verbal fluency task. The dependent variable is the inter-word intervals (IWIs) and the independent variable is the index of the IWIs.

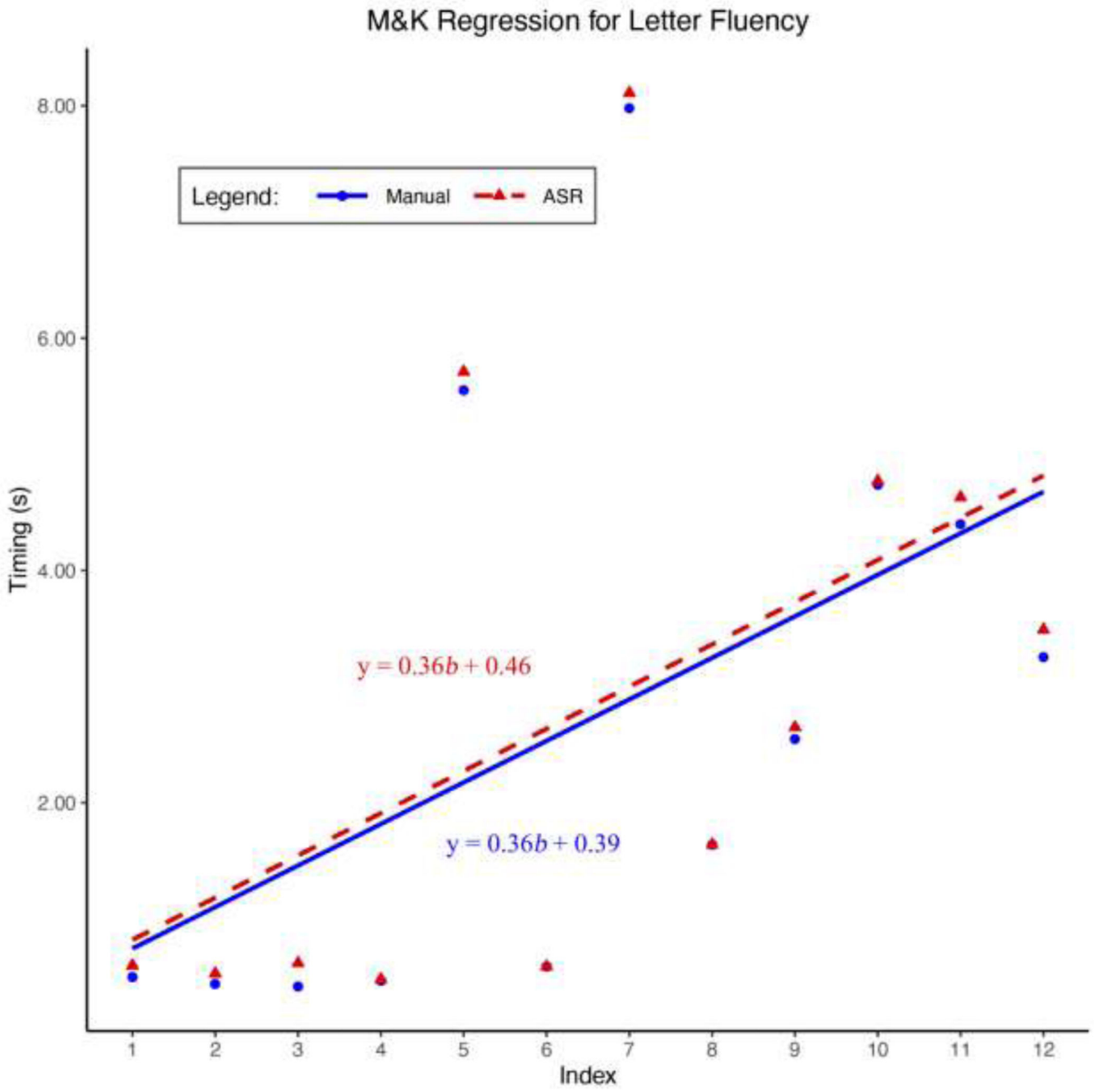

Figure 6: Mayr and Kliegl Letter F Regression

Shown are the Mayr and Kliegl linear regressions using the estimated transcriptions and manual transcriptions from a sample verbal fluency task. The dependent variable is the inter-word intervals (IWIs) and the independent variable is the index of the IWIs.

Table 1.

REGARDS participant data

	All (n=1400)	Controls (n=702)	ICI (n=698)
Age (years)	75.00 (8.68)	74.98 (8.61)	75.01 (8.76)
Sex (M:F)	643:757	323:379	320:378
Region (Non-belt:Belt:Buckle)	553:294:553	297:140:265	256:154:288
Education (< HS:HS:SC:CG+)	444:415:166:375	228:211:81:182	216:204:85:193
Race (W:B)	539:861	270:432	269:429
Animal fluency (words)	14.82 (5.10)	15.72 (5.13)	13.91 (4.91)
Letter F fluency (words)	10.05 (4.22)	10.53 (4.19)	9.57 (4.19)

Abbreviations: B, Black; CG+, college graduate or above; F, female; HS, high school; ICI, incident cognitive impairment; M, male; REGARDS, Reasons for Geographic and Racial Differences in Stroke Study; SC, some college; W, White.

Notes: Variables are shown as mean (standard deviation).

Table 2.

ASR quality scores

Metrics	Animal Fluency	Letter Fluency
VFER (%)	39.3	37.9
VFER w/PO mapping (%)	25.6	32.2
PhER (%)	23.7	19.5

Accuracies of the ASR in transcribing the audio correctly.

Abbreviations: PhER, phonemic error rate; PO, phonological overlap; VFER, verbal fluency error rate.

Table 3.

Area under the ROC curve for each classifier architecture and feature matrix

	NB	RF	SVM	Ensemble avg.	Ensemble log.
Animal fluency					
Type - Token (train)	0.9674	0.9759	0.9554	0.9708	0.9758
Type - Token (test)	0.9612 *	0.9500	0.9512	0.9623 *	0.9559
TF-IDF (train)	0.6128	0.9769	0.9558	0.9495	0.9744
TF-IDF (test)	0.5938	0.9636	0.9525 *	0.9421	0.9611
PO (train)	0.7229	0.9807	0.9528	0.9638	0.9800
PO (test)	0.7196	0.9646	0.9415	0.9475	0.9638 *
TF-IDF + PO (train)	0.7314	0.9809	0.9568	0.9651	0.9803
TF-IDF + PO (test)	0.7219	0.9649 **	0.9430	0.9495	0.9638 *
Letter F fluency					
Type - Token (train)	0.9606	0.9694	0.9417	0.9634	0.9688
Type - Token (test)	0.9544 *	0.9425	0.9424 *	0.9550	0.9393
TF-IDF (train)	0.5008	0.9687	0.9409	0.9585	0.9686
TF-IDF (test)	0.5012	0.9547	0.9404	0.9504	0.9531
PO (train)	0.8764	0.9729	0.9479	0.9672	0.9726
PO (test)	0.8821	0.9576	0.9357	0.9552 *	0.9559
TF-IDF + PO (train)	0.6649	0.9732	0.9500	0.9632	0.9730
TF-IDF + PO (test)	0.6652	0.9579 **	0.9398	0.9516	0.9568 *

NB=naïve Bayes, PO=phonological overlap, RF=random forest, SVM=support vector machine, TF-IDF=term frequency-inverse document frequency.

* Indicates the highest test set AUC for each classifier/ensemble in each verbal fluency task.

** Indicates the highest test set AUC among all classifiers/ensembles in each verbal fluency task.

Table 4.

Animal Fluency Scores

	Correlation	Manual Transcription	ASR Transcription	Unsigned diff.	Signed diff.
Repetitions	0.81	1.01 (1.25)	0.84 (1.16)	0.42 (0.65)	0.17 (0.75)
Raw score	0.91	14.94 (5.10)	14.61 (4.95)	1.56 (1.53)	0.33 (2.16)
Epoch 1	0.87	5.59 (1.85)	5.16 (1.91)	0.66 (0.83)	0.43 (0.97)
Epoch 2	0.90	3.25 (1.79)	3.10 (1.72)	0.46 (0.65)	0.15 (0.78)
Epoch 3	0.86	1.94 (1.52)	1.96 (1.41)	0.37 (0.69)	−0.02 (0.79)
Epoch 4	0.89	1.49 (1.42)	1.54 (1.32)	0.34 (0.57)	−0.05 (0.66)
Epoch 5	0.81	1.47 (1.33)	1.55 (1.35)	0.47 (0.68)	−0.08 (0.82)
Epoch 6	0.83	1.13 (1.24)	1.22 (1.29)	0.32 (0.67)	−0.09 (0.74)
c parameter *	0.72	0.06 (0.08)	0.05 (0.06)	0.44 (0.66)	−0.11 (0.79)
m parameter *	0.80	15.92 (10.12)	16.50 (10.83)	0.19 (0.89)	0.01 (0.91)
Number of switches	0.79	5.25 (2.31)	5.10 (2.05)	1.01 (1.03)	0.15 (1.43)
Max chain length	0.64	3.85 (1.83)	3.76 (1.69)	0.89 (1.21)	0.08 (1.50)
Mean chain length	0.57	1.48 (0.61)	1.44 (0.59)	0.40 (0.38)	0.04 (0.56)
M&K *b* (intercept)	0.26	−0.19 (1.84)	0.39 (3.18)	1.57 (2.87)	−0.58 (3.23)
M&K *s* (slope)	0.29	0.55 (1.33)	0.49 (0.95)	0.45 (1.32)	0.06 (1.39)
Speed score	0.91	7.84 (3.33)	7.69 (3.36)	0.98 (1.00)	0.15 (1.40)
Speed edge	0.87	5.48 (2.50)	5.60 (2.63)	0.92 (0.92)	−0.12 (1.30)
Speed switch	0.70	2.55 (1.32)	2.18 (1.06)	0.75 (0.69)	0.37 (0.95)

The data presented is from the test set (N=204). Pearson correlation was used and all other data is shown as mean (standard deviation). The unsigned (i.e., absolute value) and signed difference is presented using the actual data minus the estimated data, showing the tendency of estimated values to be larger than their counterparts.

* We used Spearman correlation for the c and m parameters and list their median values rather than mean and standard deviation. Signed and unsigned differences were calculated using z-transformation.

Table 5.

Letter Fluency Scores

	Correlation	Manual Transcription	ASR Transcription	Unsigned diff.	Signed diff.
Repetitions	0.77	0.91 (1.26)	1.05 (1.35)	0.50 (0.74)	−0.14 (0.88)
Raw score	0.88	10.07 (3.99)	10.36 (4.15)	1.46 (1.41)	−0.29 (2.01)
Epoch 1	0.87	3.50 (1.64)	3.36 (1.62)	0.48 (0.71)	0.14 (0.84)
Epoch 2	0.85	2.12 (1.33)	2.10 (1.32)	0.41 (0.60)	0.01 (0.72)
Epoch 3	0.83	1.50 (1.20)	1.58 (1.19)	0.36 (0.61)	−0.08 (0.70)
Epoch 4	0.79	1.14 (1.04)	1.30 (1.14)	0.37 (0.62)	−0.17 (0.71)
Epoch 5	0.84	0.99 (0.98)	1.07 (1.06)	0.25 (0.53)	−0.08 (0.58)
Epoch 6	0.84	0.79 (0.88)	0.86 (0.93)	0.24 (0.47)	−0.07 (0.52)
c parameter *	0.82	0.06 (0.08)	0.05 (0.07)	0.49 (0.71)	−0.12 (0.86)
m parameter *	0.78	12.56 (9.80)	12.82 (10.93)	0.39 (0.85)	0.02 (0.94)
Number of switches	0.77	3.72 (1.66)	3.79 (1.80)	0.80 (0.86)	−0.08 (1.17)
Max chain length	0.77	2.91 (1.50)	2.84 (1.48)	0.58 (0.83)	0.07 (1.01)
Mean chain length	0.57	1.22 (0.60)	1.21 (0.65)	0.37 (0.44)	0.01 (0.58)
M&K *b* (intercept)	0.74	0.81 (4.83)	1.16 (5.72)	2.18 (3.21)	−0.35 (3.87)
M&K *s* (slope)	0.66	0.87 (2.09)	0.98 (3.51)	1.00 (2.44)	−0.10 (2.64)
Speed score	0.86	4.95 (2.62)	4.41 (2.39)	1.09 (0.94)	0.54 (1.34)
Speed edge	0.80	3.60 (1.89)	3.22 (1.73)	0.90 (0.83)	0.38 (1.16)
Speed switch	0.58	1.64 (0.79)	1.56 (0.70)	0.50 (0.49)	0.08 (0.69)

The data presented is from the test set (N=204). Pearson correlation was used and all other data is shown as mean (standard deviation). The unsigned (i.e., absolute value) and signed difference is presented using the actual data minus the estimated data, showing the tendency of estimated values to be larger than their counterparts.

* We used Spearman correlation for the c and m parameters and list their median values rather than mean and standard deviation. Signed and unsigned differences were calculated using z-transformation.

Table 6.

Performance of models (N=1273, cases=634, controls=639)

	WAIC	Weight	AUC	Sens	Spec	NPV	PPV	F1
Animal Fluency								
SD timings	1586.29	0.955	**0.625**	**0.631**	**0.585**	**0.656**	**0.558**	**0.592**
SD timings, ASR	1592.42	0.045	**0.621**	**0.644**	**0.571**	**0.659**	**0.555**	**0.596**
Letter F Fluency								
Mayr scores	1757.24	0.991	**0.591**	**0.552**	**0.579**	**0.566**	**0.565**	**0.559**
Mayr scores, ASR	1766.71	0.009	**0.580**	**0.539**	**0.606**	**0.570**	**0.576**	**0.557**
Combined Model								
SD timings + Mayr	1584.08	0.931	**0.632**	**0.599**	**0.637**	**0.656**	**0.578**	**0.588**
SD timings + Mayr, ASR	1589.29	0.069	**0.630**	**0.606**	**0.621**	**0.655**	**0.571**	**0.588**

ASR = scores estimated from transcriptions derived directly from automatic speech recognition; AUC = area under the ROC curve; F1 = F1 measure (harmonic mean of PPV and Sens); NPV = negative predictive value; PPV = positive predictive value; SD = slope difference; Sens = sensitivity; Spec = specificity; WAIC = Widely applicable information criterion.