Statistical Analysis Plan for the SpeechMate Study

Birtan Demirel[1, 2], Timothy Denison[1, 2]

[1]Institute of Biomedical Engineering, Department of Engineering Sciences, University of Oxford, Oxford, United Kingdom

[2]Brain Network Dynamics Unit, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, United Kingdom

[*]Correspondence: birtan.demirel@eng.ox.ac.uk

1. Descriptive analyses

Participant characteristics will be summarised by group. Continuous variables will be summarised using mean, standard deviation, median, interquartile range, minimum, and maximum where appropriate. Categorical variables will be summarised using counts and percentages. Outcome distributions will be inspected visually before inferential testing.

2. Experiment 1: Presentation Mode Analysis

For adults who stutter, the main within participant comparison will test whether SpeechMate Presentation Mode is associated with a lower percentage of disfluent syllables than baseline Teleprompter Mode during structured public speaking. The primary analysis will use a repeated measures model with condition as a fixed effect and participant as a random effect. If the model is not appropriate because of sample size or distributional features, the paired comparison will be analysed using paired t-tests or Wilcoxon signed rank tests as appropriate.

Typically fluent adults will complete the same public speaking conditions, with and without SpeechMate Presentation Mode. Their data will be analysed using the same within participant framework where appropriate, and will also provide a reference range for speech measures and delivery ratings in fluent speakers.

For adults with apraxia of speech, the original and AI-simplified text conditions will be compared descriptively. The main apraxia relevant outcome will be articulatory accuracy together with percentage of disfluent syllables.

Secondary outcomes for the laboratory tasks include speech naturalness, presentation duration, and Stuttering Severity Instrument, Fourth Edition (SSI-4) scores. These outcomes will be summarised descriptively and analysed using the same within participant model structure, with condition as a fixed effect and participant as a random effect.

3. Experiment 2: Conversational AI Mode Analysis

Experiment 2 will use the same primary outcome and analysis framework as Experiment 1. Baseline conversational blocks without SpeechMate guidance will be compared with assisted conversational trials using Conversational AI Mode. Where sufficient data are available, scenario will be included as

an additional fixed effect. Scenario order and condition order will be considered in descriptive checks and sensitivity analyses if needed.

For the exploratory apraxia subgroup, the visual-only Conversational AI Mode will be evaluated descriptively. Consistent with Experiment 1, articulatory accuracy, usability, comfort, and speaking effort will be summarised for the assisted and baseline blocks.

4. Between group analyses

Between group analyses will compare changes in percentage of disfluent syllables, willingness to speak in public, and scores from standardised tests across groups where data are available. Models will include group as a fixed effect and participant as a random effect, with group by condition or group by timepoint interactions reported where appropriate. These analyses will be exploratory and will be interpreted alongside descriptive summaries of whether SpeechMate-supported speech in adults who stutter approaches the range observed in typically fluent speakers, and whether changes in self-reported speech experience differ between groups.

5. Follow-up analyses

Longitudinal secondary outcomes, including the Overall Assessment of the Speaker's Experience of Stuttering (OASES), the Fear of Negative Evaluation scale (FNE), and willingness to give public presentations, will be measured at baseline and after the one-month follow-up. Where paired baseline and follow-up data are available, these outcomes will be compared using a paired t-test or Wilcoxon signed-rank test, as appropriate. Outside laboratory use over one month will also be summarised descriptively using the SpeechMate follow-up logs. Logged variables will include speaking context, perceived benefit, and any problems.

6. Assumption checks and sensitivity analyses

Normality of residuals will be assessed using Shapiro Wilk tests and visual inspection. If distributions are clearly non-normal, sensitivity analyses will use nonparametric repeated measures tests, including Friedman tests across conditions and Wilcoxon signed-rank tests for planned pairwise comparisons. Planned pairwise comparisons will be corrected for multiple comparisons using a Holm approach where multiple related tests are conducted. Given the proof-of-concept nature of the study, unadjusted estimates and confidence intervals will also be reported to support interpretation.

To evaluate potential carryover or period effects inherent in the within-participant crossover design, preliminary models for both Experiment 1 and Experiment 2 will include condition sequence (the order of condition presentation) as an additional fixed effect. If significant sequence interactions are observed, the primary analysis will be adjusted accordingly, and results will be reported with appropriate caveats regarding the crossover comparisons.

7. Missing data

Analyses will use available data for each outcome. The number of participants contributing data to each analysis will be reported. If audio or video quality prevents reliable scoring of a block, that block will be excluded from analyses requiring that recording, while other available outcomes for that participant may still be used. Reasons for missing data or excluded blocks will be summarised where known.

8. Rater reliability

Independent raters blinded to condition and study aims will score disfluency measures and speech naturalness. Inter-rater reliability will be quantified using an intraclass correlation coefficient. The primary rater will repeat scoring for a predefined subset of recordings after a minimum interval of one week to assess intra-rater reliability. Where two ratings are available, analyses will use the mean score unless a prespecified adjudication procedure is required because of large discrepancies.

9. Statistical reporting

Analyses will report effect estimates, confidence intervals, exact p values where applicable, and descriptive plots or tables. Statistical significance will be assessed at a two-sided alpha of 0.05, but findings will be interpreted in the context of the exploratory proof of concept design. Any deviations from this statistical analysis plan will be documented and justified in the final study report or manuscript.