A Novel Speech Assessment Protocol for Measuring Emotional Expression: Preliminary Findings from a Feasibility Study

Felix Menne[1], Felix Dörr[1], Johannes Tröger[1], Alexandra König[1,2,3], Simona Schäfer[1], Nick Worm[4,5], Julia Koch[4,5,6],Julia Schräder[4,5,6],Lisa Wagels[4,5,6]

[1]ki:elements GmbH, Saarbrücken, Germany; [2]Cobtek (Cognition-Behaviour- Technology) Lab, University Côte d'azur, Nice, France; [3]Université Côte d'Azur, Centre Hospitalier et Universitaire, Clinique Gériatrique du Cerveau et du Mouvement, Centre Mémoire de Ressources et de Recherche, Nice, France; [4]Department of Psychiatry, Psychotherapy and Psychosomatics, Faculty of Medicine, RWTH Aachen, Germany; [5]Institute of Neuroscience and Medicine: JARA-Institute Brain Structure Function Relationship (INM 10), Research Center Jülich, Jülich, Germany, [6]Center for Computational Life Science, RWTH Aachen University

Methodological Issue

Traditional affective and cognitive assessments rely on self-report and structured tests, which are limited by subjective bias and low ecological validity. Speech-based measures offer more objective, continuous indicators of emotional and cognitive states, but it remains challenging to design paradigms that balance standardization with natural expression. Many existing tasks use autobiographical recall, which can elicit sensitive personal content and raises concerns about comfort, privacy, and data protection when speech is recorded and stored. This study introduces a semi-standardized, multi-task speech paradigm using standardized fictional prompts to elicit emotional speech without autobiographical disclosure, and evaluates its feasibility, data quality, and clinical relevance.

Background/Aims

The primary objective was to evaluate the feasibility and data characteristics of the speech paradigm administered in person.

Specifically, we assessed and hypothesized: **(1)** the proportion of participants completing all four speech tasks; we hypothesized that >75% of participants would be able to complete all tasks, **(2)** speech sample length; we hypothesized that participants would be able to produce >30s of free speech across tasks, **(3)** number of pauses and their duration; we hypothesized that these variables would not differ between task types, **(4)** correspondence between linguistic sentiment and prompt valence; we hypothesize that each prompt's valence elicits a distinct response valence, e.g., negative word ratios in answers to negative prompts differ significantly from those in responses to positive, neutral, or ambiguous prompts, **(5)** participant evaluation of task clarity, comfort, and engagement; we hypothesized an average rating of ≥4 on a scale from 1-5.

Disclosure

FM, FD, JT, AK, and SS are employed by the speech biomarker company ki:elements. JT holds shares in ki:elements. The remaining authors have nothing to disclose.

Methods

The study aims to recruit 80 participants; data from N=27 are presented. Each participant completed four storytelling tasks designed to elicit positive, negative, neutral, or ambiguous tones. For each sentiment, 20 prompts were generated by ChatGPT-4o and analyzed for sentiment; the top two were selected. Participants created short stories based on these prompts, and speech samples were recorded, transcribed, and analyzed for acoustic and linguistic features (e.g., duration, pause metrics, lexical sentiment). For all features we computed within subjects ANOVAs using the prompt valence as the within subject variable. To test aim (4) in particular, for each lexical feature, a planned within-subject contrast was defined as the difference between the target condition (e.g., negative word ratio for negative prompt) and the mean of the remaining conditions and tested against zero using one-sample t-tests.

Table 1: Speech production and lexical feature values across four prompt conditions (positive, negative, neutral, ambiguous). Features include sample length (word count, duration), pausing behavior (mean pause duration, number of pauses), and lexical characteristics (word ambiguity score; positive, negative, and neutral word ratios). Values are reported as mean ± SD. Condition effects were tested using repeated-measures ANOVAs; F, p, and η² are reported for each feature.

	Positive	Negative	Neutral	Ambiguous	F	*p*	η2
Word count	82.04 ± 64.55	87.09 ± 47.37	97.43 ± 50.71	83.83 ± 58.83	1.16	0.33	0.05
Duration (s)	55.95 ± 36.34	65.83 ± 23.83	66.73 ± 27.06	52.36 ± 25.63	3.17	0.03	0.13
Mean pause duration (s)	0.84 ± 0.45	1.32 ± 1.39	1.06 ± 0.50	0.81 ± 0.46	3.22	0.03	0.15
Number of pauses	16.04 ± 13.58	16.96 ± 10.48	16.91 ± 10.48	14.83 ± 9.93	0.51	0.68	0.02
Word ambiguity score	0.32 ± 0.13	0.50 ± 0.10	0.50 ± 0.11	0.39 ± 0.16	10.83	<0.001	0.33
Positive word ratio	0.59 ± 0.15	0.54 ± 0.07	0.53 ± 0.06	0.48 ± 0.09	4.48	0.006	0.17
Negative word ratio	0.16 ± 0.07	0.26 ± 0.07	0.25 ± 0.06	0.19 ± 0.08	10.35	<0.001	0.32
Neutral word ratio	0.21 ± 0.07	0.20 ± 0.05	0.22 ± 0.06	0.32 ± 0.07	21.11	<0.001	0.49

Results

A total of 27 participants (48.15% female; age 24.81 ± 2.91 years; years of education 14.67 ± 1.52) were included in the current analyses. Feasibility was high, with 23 of 27 participants (85.2%) completing all tasks. Story lengths ranged from 52–67 seconds and 82–97 words. Pause duration differed between conditions (0.81–1.32 s, p<0.05), whereas pause count did not (15–17 pauses, p=0.68). Responses to the negative prompt showed higher negative word ratios than responses to other prompts (p<0.02), while other prompts did not show significant effects (p=0.051–0.2). Usability ratings (n=17) averaged 4.53 ± 0.80.

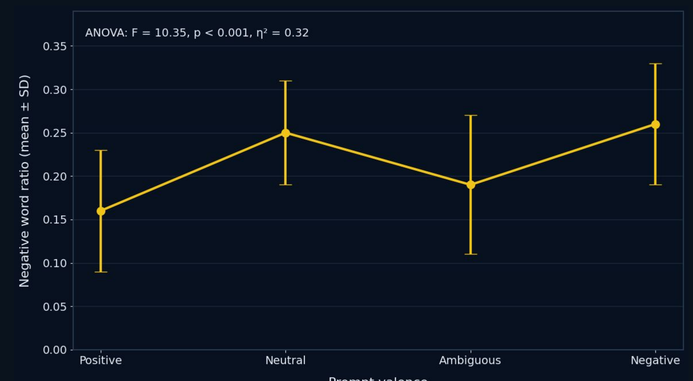

Figure 1: Negative word ratio across prompt valence conditions (positive, neutral, ambiguous, negative). Points show condition means and error bars indicate ± SD. A significant effect of prompt valence was observed (repeated-measures ANOVA: F = 10.35, p < 0.001, η² = 0.32)

Conclusion

Findings indicate that the semi-standardized paradigm is feasible, engaging, and capable of producing high-quality speech samples. High completion rates and consistent story lengths support usability. Differences in pause duration but not frequency suggest sensitivity to task variation without compromising comparability. Valence alignment emerged only for the negative prompt, indicating detectable but asymmetric emotional correspondence. Overall, the paradigm shows promise for assessing affective expression, with opportunities to refine prompts to improve differentiation across conditions.