SPEAKapp – Remote monitoring of language production to predict cognitive functioning

Chiara Barattieri di San Pietro[1], Valentina Simonetti[1], Cristina Crocamo[2], Maria Bulgheroni[1]

[1]Ab.Acus s.r.l., Milan, Italy

[2]Department of Medicine and Surgery, University of Milano-Bicocca, Italy

https://doi.org/10.36505/ExLing-2021/12/0009/000482

Abstract

Language production and comprehension can provide a useful perspective into an individual's mental health and cognitive abilities. SPEAKapp is a mobile application designed to deliver and analyze speech and language data for clinical and research purposes. It implements pre- and post-processing techniques based on Natural Language Processing (NLP) and Distributional Semantic Models (DSM) of language. The first functional prototype was tested for accuracy of data acquisition and elaboration, as well as for usability and acceptability in a pilot sample of fragile users. SPEAKapp showed good accuracy and replicability of results, and participants felt comfortable using the application. Further developments of the application are presented.

Keywords: language, NLP, mobile app, DSM, neuropsychology.

Introduction

Speech and language assessment is a powerful yet unobtrusive tool that supports clinical assessment in various conditions implicating cognitive and affective symptoms. The current neuropsychological practice relies on standard paper-and-pencil batteries, requiring verbatim transcriptions and manual scoring. This is time-consuming and represents a barrier to frequent monitoring and remote assessment.

Technically, verbal responses are a stream of data suited to be processed with modern speech technologies. On the one hand, available automatic speech recognition software has reached remarkable accuracy. One the other, speech content can be modelled by relying on DSMs, which are automatic data-driven models of semantic representations that represent word meanings as numerical vectors in multi-dimensional spaces (Landauer & Dumais, 1997), and that can be used to simulate the structure of conceptual knowledge implied in the performance of semantic tasks (Mandera et al., 2017). However, to build upon these technologies and address actual clinical and research needs, it is pivotal to bridge the gap between end-users and clinicians on the one hand and developers and researchers on the other.

To do so, we developed "SPEAKapp", a system based on a mobile application to deliver and analyze speech and language data for clinical and research purposes. Precisely, SPEAKapp can: i) deliver standard neuropsychological tests and questionnaires; ii) collect audio responses; iii) perform speech-to-text transcription and noise removal, iv) score the results using digitalized deterministic approaches and semantic analysis; and v) store and manage output data. SPEAKapp frontend is implemented using the Flutter framework, whereas the backend is hosted on a cloud server. SPEAKapp relies on a commercial speech-to-text service (https://cloud.google.com/speech-to-text), as well as property algorithms to extract related meaningful semantic features implemented in Python. The system is GDPR compliant.

The main goal of the present study was to test: i) the reliability of the SPEAKapp system in terms of accuracy of automatic transcriptions; ii) the portability of the logic from a research environment to the final product environment, and iii) the usability and acceptability of the system in a pilot sample of potential clinical users.

Methodology

A sample of 23 participants, Italian native speakers, among which 8 users of community mental health services, was recruited via personal referral. A categorical Verbal Fluency (VF) task was delivered through the app. A VF task is a standard neuropsychological test used to assess lexical retrieval: participants are asked to produce as many words as possible in a given semantic category (i.e., "animals") within a time limit (60 sec).

Automatic transcriptions were inspected for accuracy. Accuracy was calculated as TP + TN / (TP + TN + FP + FN), where TP is the number of valid tokens correctly identified (True Positive), TN is the number of tokens correctly ignored as irrelevant (True Negative), FP is the number of tokens incorrectly transcribed (False Positive), and FN is the number of tokens not recorded although relevant (False Negative).

The results of the semantic analysis computed by the app were compared against the results of the original algorithms (Barattieri di San Pietro et al., 2020) in terms of the size of semantic clusters (the number of consecutive words produced that share similar properties) and the number of switches (i.e., the total number of transitions between these groups – Troyer et al., 1997) as computed from a word2vec (Mikolov et al., 2013) semantic space.

A System Usability Scale (SUS – Brooke, 1996) questionnaire was administered to the subgroup of participants who were users of community mental health services at the end of the data acquisition phase to test the acceptability and usability of the system.

Results

The total number of spoken output (including sound fragments, repetitions, confabulations, conjunctions, false starts, and words) was manually identified (N = 782). SPEAKapp recorded and transcribed N = 613 tokens. Compared to the manual transcription for clinical purposes (N = 560 words), SPEAKapp recorded and transcribed N = 68 irrelevant tokens (0.12%), and failed to transcribe N = 15 relevant tokens (0.03%). The resulting accuracy for research purposes of SPEAKapp was overall 89.71%.

On average, participants produced 23.13 words each (standard deviation SD = 8.34). The mean number of switches was 9.61 (SD = 6.01). The pooled mean of cluster size was M = 3.05 words. The analysis was carried out twice, both with the original algorithms and the Python-embedded logic in SPEAKapp, and yielded identical results.

Analysis of the SUS questionnaire results revealed that participants felt comfortable using the application, whose functionalities were considered well integrated. Although not entirely autonomous in the use of the app, users felt that with a little initial training, they were able to learn what was needed to perform the tests. Overall results indicated that the users would feel comfortable using SPEAKapp again.

Discussion

The present work aimed to test and validate the first functional prototype of SPEAKapp, a mobile application for language assessment that implements both standard deterministic approaches to test scoring as well as a set of novel indexes based on NLP and DSM. Results showed that the accuracy was adequate for clinical purposes and that data loss compared to manual transcription was nearly negligible. Results of the semantic analysis were in line with the expected, showing the successful translation to a commercial environment. Finally, the use of the application was considered practicable by participants.

To accommodate the needs of clinicians and researchers, the updated version of the app implements a comprehensive range of neuropsychological tests, such as ad-hoc versions of a prose recall test, a picture-naming test, a test with repetition of complex sentences, and an automatic series test. Tests are available in multiple languages. Given the growing interest in the analysis of speech acoustic analysis for clinical purposes, a PRAAT library (Boersma & Weenink, 2021) was implemented to extract fundamental frequency (F0), cycle-to-cycle perturbation measures of pitch (jitter), and amplitude (shimmer) as well as the harmonic-to-noise ratio (HNR). A designated web-based dashboard will enable clinicians to interact with the system for data entry and visualization. Collected data will feed a set of Machine Learning models on clinical outcomes based on changes over time of both standard and experimental indexes of language production to predict clinical visit annotations (target).

Beyond being a practical tool for data acquisition and transcription, the adoption of additional linguistic measures can measure subtle differences in language use. The integration of novel indexes based on verbal performance with standardized neuropsychological measures might lead to novel insights into mental health conditions, as well as to the identification of light and reliable indexes of cognitive functioning. In the long term, identifying a valid marker of treatment efficacy would support clinical research and innovation, facilitating the evaluation of new drugs' efficacy and the design of targeted approaches.

Acknowledgements

SPEAKapp has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 857223.

References

Barattieri di San Pietro, C., de Girolamo, G. C., Luzzatti, C., Marelli, M. 2020. Mind your models! Distributional semantic models for the analysis of verbal fluency tasks in Schizophrenia Spectrum Disorders. The 26th Architectures and Mechanisms for Language Processing Conference, virtual congress.

Boersma, P., Weenink, D. 2021. Praat: doing phonetics by computer [Computer program]. http://www.praat.org/

Brooke, J. 1996. SUS-A quick and dirty usability scale. Usability evaluation in industry, 189(194), 4-7.

Landauer, T. K., Dumais, S. T. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211.

Mandera, P., Keuleers, E. & Brysbaert, M. 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. Journal of Memory and Language, 92, 57-78.

Mikolov, T., Chen, K., Corrado, G., Dean, J. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781)

Troyer, K., Moscovitch, M. Winocur, G., 1997. Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. Neuropsychology, 11(1), 138–46.