scientific data

OPEN

DATA DESCRIPTOR

Slovak database of speech affected by neurodegenerative diseases

Milan Rusko[1]✉, Róbert Sabo[1], Marián Trnka[1], Alfréd Zimmermann[2], Richard Malaschitz[2], Eugen Ružický[3], Petra Brandoburová[4,5,6], Viktória Kevická[1,7] & Matej Škorvánek[8,9]

A new Slovak speech database EWA-DB was created for research purposes aimed at early detection of neurodegenerative diseases from speech. It contains 1649 speakers performing various speech and language tasks, such as sustained vowel phonation, diadochokinesis, naming and picture description. The sample of speakers consists of individuals with Alzheimer's disease, mild cognitive impairment, Parkinson's disease, and healthy controls. In this article we describe the EWA-DB development process, the language and speech task selection, patient and healthy control recruitment, as well as the testing and recording protocol. The structure and content of the database and file formats are described in detail. We assume that the presented database could be suitable for the development of automatic systems predicting the diagnoses of Alzheimer's disease, mild cognitive impairment, and Parkinson's disease from language and speech features.

Background & Summary

Alzheimer's disease (AD), and Parkinson's disease (PD), are the two most common neurodegenerative diseases, which have been historically diagnosed based on their clinical symptoms, however, additional diagnostic tools such as imaging or biological markers are being increasingly implemented in their clinical diagnostic criteria[1,2]. While these diagnostic criteria require specific clinical expertise or availability of more sophisticated investigation methods, the need for early (even prodromal) diagnosis and screening of subjects at risk is becoming increasingly important. In this regard, mild cognitive impairment (MCI) seems to be the most relevant risk factor for development of fully manifested Alzheimer's disease, and updated research criteria for prodromal Parkinson's disease have been also published recently[3,4].

According to the World Health Organization more than 55 million people worldwide have AD or other forms of dementia and there are nearly 10 million new cases every year[5]. The patient's speech and language are commonly affected by these disorders and specific changes can be observed even in their prodromal phases[6], which can be utilized in disease diagnostics and further monitoring.

As for speech and language, Alzheimer's disease is most pronounced at the lexical-semantic, discourse-pragmatic, syntactic and phonetic levels[7]. Studies using machine learning techniques for speech analysis for MCI and AD patients show that different parameters have different weights for the diagnosis of AD and MCI and that the combination of several parameters improves the accuracy of neurodegenerative disease prediction[8]. In Parkinson's disease, motor speech problems in the sense of hypokinetic dysarthria prevail, which affect several subsystems of speech - phonation, resonance, articulation and prosody[9] and are captured by parameters such as pause duration[10] and prosody change[11]. Symptoms of PD are present not only in speech but also in language, mainly in morpho-syntactic processing, as in verb processing[12].

As summed up by Chen[13] assessment of cognitive function is typically carried out by a trained psychometrician or neuropsychologist using a battery of cognitive tests that examine various aspects of cognitive abilities, including language skills[14,15]. Neuropsychological testing is a time-consuming process that can take up to several hours. Also, in many cases repeated assessments are necessary to monitor the progression of cognitive decline. There is currently a growing interest in creating automated assessment methods that will speed up the process

[1]Institute of Informatics of the Slovak Academy of Sciences, Bratislava, Slovakia. [2]AXON PRO, Bratislava, Slovakia. [3]Faculty of Informatics, Pan European University, Bratislava, Slovakia. [4]Department of Psychology, Faculty of Arts, Comenius University, Bratislava, Slovakia. [5]MEMORY Centre, Bratislava, Slovakia. [6]2nd Department of Neurology, University Hospital, Bratislava, Slovakia. [7]Department of Communication Disorders, Faculty of Education, Comenius University, Bratislava, Slovakia. [8]Department of Neurology, Faculty of Medicine, P.J. Safarik University, Kosice, Slovakia. [9]Department of Neurology, University Hospital L. Pasteur, Kosice, Slovakia. ✉e-mail: milan.rusko@savba.sk

of early detection of people with neurodegenerative diseases based on their cognitive and especially language difficulties[15].

The existence and availability of a speech database that represents the investigated phenomena on a sufficient sample of patients with neurodegenerative diseases and healthy people is a necessary condition for research into the possibilities of using automatic speech processing and machine learning methods to predict neurodegenerative diseases from speech.

For the two most widespread neurodegenerative diseases, PD and AD, there are several types of speech databases in the world. Some of them are published and accessible in the form of corpus description, fewer of them are published in the form of recordings. Some of them are accessible on request and the smallest group are freely accessible PD[16] and AD[17] databases that can be directly used for machine learning.

Furthermore, we can divide speech databases into groups according to the type of tasks that are recorded during its creation. The most widespread types of tasks that the databases contain are for PD[18,19]: phonation of sustained vowels, reading and repetition of sentences, description of pictures or diadochokinesis tasks (e.g., rapid repetition of syllable sequence /pataka/); and in patients with AD[20,21]: picture description, fluent questions answering, object naming or narratives on the topic of daily life.

To illustrate some examples of PD databases, we can highlight several notable collections. The *New Spanish Speech Corpus Database* (*PC-Gita*)[22] contains speech recordings from 100 native Spanish speakers, equally divided between 50 individuals diagnosed with PD and 50 healthy controls matched by age and gender. The *Parkinson's Voice Initiative (PVI) Dataset*[23] was gathered through a smartphone application, capturing sustained vowel phonation from participants across seven countries. The *Parkinson's Speech with Multiple Types of Sound Recordings*[24], collected at Istanbul University, includes recordings of various speech tasks such as sustained vowel phonation, counting from 1 to 10, and reading short sentences and words. Frequency-based features were extracted from each voice sample and are provided alongside the Unified Parkinson's Disease Rating Scale (UPDRS) score of each patient. The *Michael J. Fox Foundation (MJFF) Parkinson's Progression Markers Initiative (PPMI) database*[25] includes speech samples from individuals with Parkinson's disease as well as control subjects, accompanied by demographic and clinical information. The *NeuroVoz dataset*[26], available in Castilian Spanish, offers a variety of speech tasks, including sustained vowel phonation, diadochokinesis, structured listen-and-repeat utterances, and spontaneous monologues. The *Mobile Device Voice Recordings at King's College London* (*MDVR-KCL*)[27] contains recordings of both early and advanced Parkinson's disease patients, along with healthy controls, all collected via phone calls. Speech tasks included text reading and spontaneous dialogues. The *Oxford Parkinson's Telemonitoring Dataset*[28] resulted from a six-month trial involving 42 individuals with early-stage Parkinson's disease, focusing on remote symptom progression monitoring through telemonitoring devices.

For AD databases, several key datasets are noteworthy. The *Talkbank DementiaBank's Pitt Corpus*[29] includes speech recordings from individuals diagnosed with probable or possible AD, MCI, memory impairments, vascular dementia, as well as control subjects. The *SpeechDx*[30] initiative aims to develop speech-based biomarkers for early detection of Alzheimer's disease and related dementias by building a large dataset of speech data linked to clinical data from individuals across the brain health spectrum. According to Sevcik[17], the *Alzheimer's Dementia Recognition through Spontaneous Speech* (*ADReSS*) database[31] is widely used in comparative experiments for model training. However, this database is actually a subset of the Pitt Corpus[29]. Some researchers have also utilized the *Carolinas Conversation Collection* (*CCC*)[32] in their experiments. While most AD datasets are in English, there are also recordings in Mandarin[33], Italian[34], Swedish[35], Nepali[36], French[37], and other languages.

Despite the availability of numerous PD and AD datasets, they often face challenges such as inconsistency across datasets, extremely specific purpose of use, poor quality of recordings or dataset imbalance[38,39]. Another problem is the small number of recordings in the dataset. In this case, there is not enough data for training the machine learning techniques cannot show their full potential. Databases are also language specific, and it is problematic to find different datasets which can be combined.

Language specificity creates perhaps the biggest challenge. To name an example, phonemic diversity in different languages represents a significant challenge for the evaluation of speech/language and their automatic processing through machine learning and ultimately for the final prediction of neurodegenerative diseases[40]. For instance, Anglo-Germanic languages are characterized by frequent clusters of consonants, while in Roman languages the consonant-vowel structure predominates[41]. Due to language differences and the challenges, they can generate, we find the creation of new language specific databases necessary.

Speech databases are essential for training automatic speech recognition (ASR) systems, enabling thorough testing and benchmarking to ensure these systems perform effectively across a variety of conditions and speaker variations. Additionally, text-to-speech (TTS) systems rely on extensive and diverse speech datasets, including those from individuals with neurodegenerative diseases, to produce natural-sounding synthetic speech. From a medical perspective, emerging areas of research are increasingly focusing on speech and language, particularly at the discourse level of language processing. Moreover, the development of new diagnostic tools, both traditional and digitalized, necessitates the existence of language and diagnosis-specific databases, underscoring their critical importance.

This manuscript presents a new Slovak speech database created for AD, MCI, and PD research and for development of automatic systems predicting these diagnoses from language and speech features.

Methods

As manifested in the previous section, creation of language specific databases is needed. When creating the Slovak database, we followed the current International Parkinson and Movement Disorder Society Guidelines for Speech Recording and Acoustic Analyses in Dysarthrias of Movement Disorders[40]. We also considered the language and speech testing tasks, which are used in common clinical practice and at the same time are shown

by research to be suitable materials for obtaining a speech sample for its automatic analysis. As for the participants, we considered it important to represent the most common neurodegenerative diseases, which are PD and AD/MCI, as well as the healthy population for the purposes of creating the possibility of comparison with normative performances.

Selection and compilation of speech and language tasks. Speech and language tasks that are known to be sensitive to capture early changes in the speech and language production of PD, AD and MCI patients were selected.

Sustained vowel phonation and diadochokinesis. Since a high percentage of PD patients have voice difficulties related to dysarthria, it is crucial to include tests that evaluate these symptoms. Sustained vowel phonation and sequential and alternating movement tasks, such as diadochokinesis, are among the basic voice tasks included in the voice diagnosis of patients with PD.

For the purposes of this project, following the guidelines by Rusz[40], participants were first instructed to take a deep breath and perform a sustained phonation of vowel /a/ as long and steadily until they run out of air or until the end of recording, which was set at 15 seconds. The ideal minimum duration of phonation recommended for hypokinetic dysarthria patients, including PD, is 6 seconds[41].

Diadochokinesis is a dysarthria evaluation method commonly applied in clinical practice, which tells about the ability of the maximum speed of syllable repetition using alternating or sequential movement tasks. The alternating task measures the rapid repetition of a single syllable, while the sequential task measures the rapid repetition of syllable sequences[42].

In our project, the sequential motion rate was included. Participants were instructed to take a deep breath and repeat syllable sequence /pataka/ as quickly and accurately as possible until told to stop. The instruction was to pronounce continuously, intelligibly and to speak as quickly as they can without being imprecise[43]. The duration of this test recording was set at 8 seconds to allow for at least 12 sequence repetitions performed with one breath as suggested by the above-mentioned guidelines[40].

Object and action naming. Confrontational naming relies on specific brain networks, and involves various cognitive processes such as visual recognition, semantic activation, lexical retrieval and articulation[44,45]. In the task of confrontational naming, perceptual difficulties, visual perception errors, impaired phonological and semantic access is documented in patients with AD[46]. Difficulties in picture naming are also documented in PD[47].

There are several visual confrontation naming measures available. The most often used confrontational naming test is the Boston Naming Test (BNT). It is commonly used by neuropsychologists and speech therapists to assess lexical retrieval. The original version includes 85 items as simple line drawn pictures[48] and it was later shortened to 60 items[49]. The Slovak version of the test is not available so far. An original Slovak picture naming test is available[50], containing black-and-white drawings of 30 objects and 30 actions. However, this test appears to be insufficiently sensitive to discrete language disruptions as those manifested in early PD[51].

In the design of the visual confrontation naming task for EWA-DB our own set of 30 object and 30 action photographs was created. Choosing the format for full colour was based on the work of Li *et al.*[45] which describes several reasons for coloured image preferences, such as criticism of the black-and-white line drawings in the BNT[52], possible cohort effects and cultural bias that make these drawings liable to be misperceived in non-English speaking population[53], or a questionable diagnostic validity[54]. In comparison, a meta-analysis of studies aimed at picture naming showed improvement in naming accuracy and response times when using coloured images[55]. In the study of Li *et al.*[45] the original version of the BNT and a colour version of BNT were compared. The study documented higher scores in naming accuracy and an overall better diagnostic accuracy for MCI and AD detection when the coloured version of the test was used[45].

In addition to colour, another factor affecting picture naming ability is the nature of the objects in the sense of biological and artifact items. When the ability to name biological and artifact items was observed, individuals with MCI showed lower performance on biological items compared to cognitively healthy individuals, suggesting a category-specific impairment of biological items in MCI[56–59].

Not only objects but also actions are part of the naming test created for EWA-DB. The reason is that naming processes for verbs and nouns are different. As explained by Hwang *et al.*[60], verbs and nouns belong to separate grammatical classes and the difference is also in terms of their semantic representation as nouns refer to objects, while verbs refer to actions. Nouns and verbs engage different brain areas. Lexical retrieval of nouns is engaging left temporal areas and verb retrieval is engaging left frontal areas[60,61]. This could be one of possible explanations for difficulties of verb production in PD. Several studies of PD document action naming deficits, thought to reflect the presence of frontal and prefrontal dysfunction[62], namely in the areas of pre- and post-central gyri bilaterally, left frontal operculum, left supplementary motor area[63].

In the final picture set for EWA-DB 30 nouns and 30 verbs were selected. Their selection was based on the criteria of age of acquisition, frequency, and word length. Some additional criteria for the visual form of stimuli were added:

- The number of high-frequency words is 1/3 compared to low-frequency words. The frequency is determined based on information from the Slovak National Corpus. For a word to be considered high frequency, it's occurrence in Slovak language had to be greater than 1000.
- The pictures had to be as simple as possible so that there is no doubt about what the main object or main activity of the image is. Images without a background were chosen to focus attention on the object itself.
- The pictures could not be too detailed as the pictures were planned to be presented by smartphone screens and given their age, we assumed that some users may have vision problems.

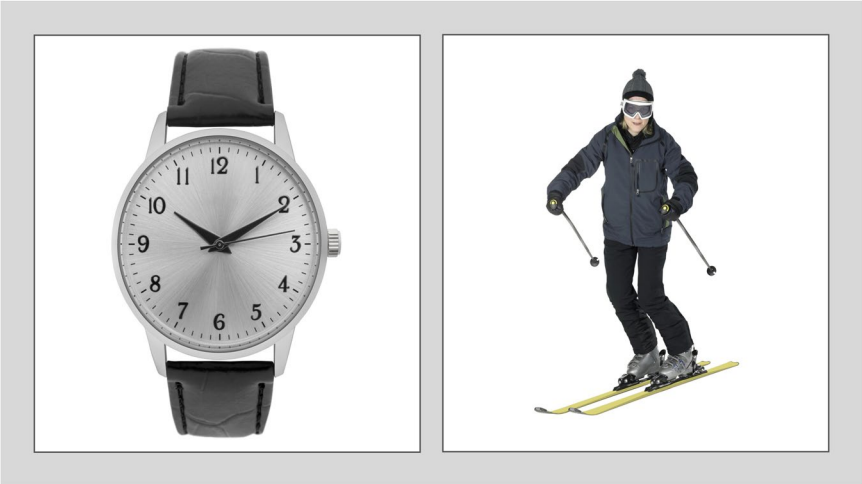

Fig. 1 Example of object and action naming items.

- The pictures had to be square or vertically oriented rectangle, so that their details can be recognized well on the smartphone screen in a vertical position.
- The pictures are intended for Slovak speakers, mainly in the age of 50–80. Objects and activities were chosen that were well known and frequently encountered by these persons during their youth and active adulthood. For example, a soccer ball would be selected for a ball image, rather than an American football ball.

An example of images used in object and action naming is given in Fig. 1.

Picture description. The analysis of spontaneous speech has been receiving more and more attention. In AD, changes in lexis, grammar, informativeness[64], cohesion and coherence are documented, while the most evident are the changes in fluency and semantics[65]. Since already in MCI and early AD disruptions of the temporal parameters of speech can be found (speech tempo, number of pauses and their length) the computerized analysis of spontaneous speech may be a promising tool for early AD detection[66]. Disruptions of spontaneous speech are also captured in PD. The spontaneous speech of patients with PD is less fluent, monotonous, and syntactically simpler with a disturbance in the informativeness present[67]. In recent years, attention has also been paid to cohesion and coherence in early PD. A higher number of incorrect connective ties and thus a lower degree of cohesion is documented[68]. Difficulties in this area are present not only in cohesion, but also in coherence, namely global coherence[69].

There are several ways to obtain a sample of spontaneous speech, such as interviewing, storytelling or picture description. Spontaneous language analysis with a picture description task is useful to detect subtle language impairments even in early stages of AD[70]. Also in PD, spontaneous speech assessment is considered to have the best ability to differentiate PD patients from healthy adults[71].

The most widely used task in neurodegenerative diseases is the Cookie Theft from the Boston Aphasia Diagnostic Examination[72]. This task has proven to be an effective tool in several neurogenic disorders, including AD and MCI[20]. However, for the purposes of our project, we encountered several difficulties in implementing this image, such as its landscape orientation not being suitable for mobile phone displays or the inscription being in English. We also wanted to remain consistent when creating images for the database, so we preferred colour images over black and white. We expect the advantages of colour images to be the same as we describe in the section on creating images for the naming task. We therefore created new images for this project, while we tried to incorporate those features of the Cookie Theft picture, which make it such an effective tool, as described by Cummings[73]. The pictures we created contain people, objects, activities and situations, the description of which can best capture the differences between healthy expression and the expression of a person with beginning AD or PD. According to Cummings[73], there are 7 areas assessed in picture description. Namely points 2 to 5 had to be considered in the design of our images by incorporating people, objects and activities containing areas and categories that are important for the assessment of our patient's speech. The areas of interest are as follows:

- **Significant features, importance of information**. The patient should present a clear distinction between essential and minor matters; the chronology of his statement should be correct in the sense of starting with the essentials.
- **Semantic categories**. AD patients show a tendency to use more general, less specific terms e.g., *woman* is more general than *mother*. The picture also must depict actions e.g., *the boy falls*.
- **Referential cohesion**. The use of deixis (*he*, *she*) in a way that it is clear which person or object the deixis represents, which is a problem for AD and PD patients.

Fig. 2 Example of a picture used in the picture description task.

- **Causal and temporal relations**. The patient should describe causalities - that the water flows out because the woman left the tap open; the children steal biscuits because they saw that the mother was not looking etc. In AD the context is often incomprehensible.
- **The language of mental state**. Cognitive mental state includes knowledge, beliefs, and assumptions. Affective mental state is represented by emotions such as happiness or anger - the events and actions in the picture can only be explained by the mental state of the persons. Typical descriptive words of the language of the mental state are for instance *wants*, *dreamy*, *careful*, *forgot about* etc. In AD they are often missing.
- **Language structure and speech**. Assessment of language and motor speech structure skills, such as word search, pauses, auxiliary sounds, replacement of words with a more general form (e.g., *that*, *stuff*, *someone*, *person*), neologisms and other phonological errors.
- **General knowledge and perception**. Some defects in general knowledge and perception, such as forgetting that one has already described something and describes one scene several times or giving a description of only one of the sides of the image (right or left). However, this occurs rather in patients with severe dementia[58].

Based on the described criteria, five pictures were created for the purposes of the EWA-DB – two simple black-and-white pictures and three complex coloured pictures. An example of these pictures is shown in Fig. 2.

Recruitment of participants. Since the goal of this project was a new Slovak speech database created for AD, MCI and PD research, the recruitment of such patients along with healthy controls took place.

To recruit a large number of participants from all regions of Slovakia, we reported on the ongoing project in newspapers, on the radio, and on television. We distributed leaflets mainly to retirement homes and general practitioners' clinics. An association of retirement companies provided a database of 140 facilities that were approached with a request for cooperation. From the approached organizations and contacts, the interested parties announced themselves via e-mail or telephone. Subsequently, the interested party was sent an informative article about the details of the research process and a personal meeting was arranged. MCI and AD participants were mainly recruited from the MEMORY Centre - a specialized centre for diagnosis, treatment, and education in the field of memory disorders and dementia. PD participants were mainly recruited from the Department of Neurology and Centre for Rare Movement Disorders of the L. Pasteur University Hospital. So, the group of positively diagnosed patients consists of people who have already been diagnosed with MCI, AD, or PD during a medical examination.

During the recruitment the following methods and examinations were used.

The anamnestic questionnaire contains items for obtaining basic descriptive characteristics (gender, age, education, and lifestyle factors). Although the collection of samples of speech recordings is not quota-based, the aim is a heterogeneous proportional representation in relation to the monitored variables.

The subtest Similarities from the WAIS-III[74] is considered one of the best instruments for measuring the index of the verbal comprehension factor, in which crystallized intelligence (learned procedures and knowledge) is significantly involved[75–77]. Premorbid intelligence is a concept through which the impact of neurological damage on cognitive performance is evaluated as an estimate of the basic, "premorbid" global performance before the onset of damage[78]. The total score in the subtest is an indicator of the current level of the examinee's abstraction ability and overall verbal comprehension. The test is also recommended as a suitable tool for the evaluation of semantics and language from the available tests adapted in our region.

Montreal Cognitive Assessment (MoCA)[79] is a screening tool aimed at assessing the global cognitive status by including the assessment cognitive domains such as executive functions, visuospatial functions, memory, attention, speech, orientation in time and space. The scoring ranges from 0–30 points. The test itself states a cut off score of 26 points. However, in older adults, the average performance ranges from 26 ± 3 points and is also dependent on education[80].

Barthel's Index for Activities of Daily Living[81] is a tool measuring the extent to which someone can function independently in their activities of daily living (e. g. dressing, feeding, bathing etc.). In our research the self-report version of this tool was administered.

Geriatric Depression Scale (GDS)[82] is a questionnaire assessing the presence of depressive symptoms in older age. It assesses symptoms present in the current experience. The patient's task is to select the answer from 15 questions that best describes how they have felt in the past week. In our research the Slovak adaptation of this questionnaire was administered[83].

Generalized Anxiety Disorder 7 (GAD-7)[84] is a questionnaire that measures the severity of symptoms associated with anxiety or anxiousness. Patients indicate how often they have been bothered by any of the difficulties formulated in the 7 questions over the past two weeks. When used as a screening tool, further evaluation is recommended when a score of 10 or higher is obtained. In our research, the Slovak adaptation of the questionnaire was administered[83].

Inclusion and exclusion criteria. When creating the database and recruitment of participants, we followed these **inclusion and exclusion criteria**:

Mild cognitive impairment (MCI). Classification criteria according to the diagnostic criteria of Albert *et al.*[85], age 50–90 years, MoCA score 25-23, preserved activities of daily living measured by Barthel's Index for Activities of Daily Living[81], score in the Geriatric Depression Scale (GDS)[82] $\leq$9 b, General Anxiety Disorder-7 (GAD-7)[84] $\leq$9 b. Participants had to be diagnosed as patients with MCI by a psychiatrist or a neurologist according to the Albert *et al.*[85] criteria for a diagnosis of MCI. Exclusion criteria for the MCI group were as follows: (1) history of or current psychiatric disorder; (2) history or neurological evidence of stroke, head injury, or neurodegenerative disorders that are known to influence cognitive functioning; and (3) on medication for depression and/or Alzheimer's disease.

AD. Diagnosis established according to the criteria of the International Classification of Diseases (ICD-10) or the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) and in all stages of the disease confirmed by a specialist (psychiatrist or neurologist), age 50–90 years, MoCA score between 22-18, capability to give informed consent.

PD. Diagnosis of manifest PD was based on the International Parkinson and Movement Disorder Society (MDS) clinical diagnostic criteria[2]. Prodromal PD cases were recruited from the PARCAS and PDBIOM cohorts based on the updated MDS criteria for prodromal Parkinson's disease[4] as described previously[86]. Patients were recruited in prodromal or early stage of the disease (Hoehn & Yahr stage I or II) and duration of the disease less than 10 years, MoCA score $>$ 20 points, age over 18.

All participants also met the inclusion criteria: normal or compensated vision and hearing, being a native speaker of Slovak language, no clinical history of head injury or psychosis, no medical record of drugs or alcohol consumption, not being under pharmacological treatment affecting cognitive functions, absence of disorders with expected impact on language and speech.

Exclusion criteria for healthy and clinical groups were the self-reporting of a previous stroke, brain tumours or psychiatric disorders such as bipolar disorder or schizophrenia, current or past alcohol or drug abuse history, or under-corrected auditory or vision difficulties.

The recruited participants were invited to various recording sites across Slovakia and underwent the same recording procedure administered by trained staff (administrators). The team of administrators consisted of healthcare workers, such as psychologists, doctors, nurses, and trained staff or students (mainly students of psychology or speech and language pathology).

Testing and recording procedure. The testing and recording procedure were performed in the following steps. Participants signed a written informed consent prior to participating in the research, including information about personal data processing, providing health status information and sound recordings publishing. The study was undergone in accordance with the Declaration of Helsinki, including the ethics committee approval of the Bratislava self-governing region (committee number 03187/2021/HF). The Ethics Committee approved to conduct the biomedical study and share the data to the extent of informed consent (i.e. using the data for scientific and commercial research and provision of data to third parties).

Following that, the anamnestic questionnaire and measures for the inclusion criteria were administered (see above for details). Finally, the sound recording protocol was administered. It consisted of speech and language tasks: sustained vowel phonation, diadochokinesis, object and action naming and picture description, the designs of which were described in detail in the section above.

Data collection and processing. Commonly available mobile phones were used for data collection. Both Apple and Android devices have been used. At the beginning of the project, older phone models were also tested, and the recording software was modified to use only mono recording on all devices, without the use of compression with 16 bit depth and sampling rate of 16 kHz. Audio files were saved in Microsoft waveform format with Pulse Code Modulation.

The recording procedure started with a calibration phase during which the usable screen size was determined, the functionality of the recording and the reliability of automatic speech recognition on the voice of the participant were verified.

A smartphone application was designed that covers the entire process of collecting data from participants:

- Anamnestic questionnaire. Basic information about the participant (age, gender, education, lifestyle factors) is inserted by the administrator. Sensitive data (first name, last name, contact details) are protected by the GDPR law and are not processed electronically. They are collected in writing and archived together with the data processing consent in accordance with the GDPR law.
- Measures for the inclusion criteria (see Inclusion and exclusion criteria).
- Sound recordings of sustained vowel phonation and diadochokinesis.
- Sound recordings of picture naming. The application sequentially displays 60 pictures (30 objects and 30 actions). Participants are instructed to name the object or action using one word. Moving from one picture to another is done by pressing a button. As we discovered at the beginning of the project, people tend to move to the next image too soon, press the button and end the recording before finishing the whole word. Therefore, the application stops recording with a short delay. The standard recording time of one image is 2-3 seconds.
- Sound recordings of picture descriptions. The application presents 2 simple black and white pictures and 3 complex color pictures. Participants were asked to describe these pictures in detail. The recording time was set to 30 seconds for simple pictures and 90 seconds for the complex ones.
- Informed consent for data processing and publication.

The recording process was conducted with the help of trained personnel who tested the participants to verify inclusion criteria and assisted in administering the tests using the mobile application. This ensured a smooth process and that all participants received proper guidance during the speech task recordings.

A server application was created to collect and process data from mobile phones used in our project. All data from the mobile phones, data supplemented with the date of testing, name of the administrator, and technical information about the version of the application and the mobile phone itself were uploaded to a MySQL database[87].

A total of 120 answers to questions from the questionnaires used in the inclusion criteria and 65 audio recordings per participant, are stored in a database. It is possible and recommended to record when the mobile is disconnected from the internet and in airplane mode so that the test is not interrupted. The application sends the data to the server when it is connected to the internet. The server uses a MySQL database[87] to store metadata. The audio data is stored in the Minio[88] database. Due to the different devices that can be used, recordings are stored in PCM Microsoft wav format with 16 kHz sampling and 16 bit resolution. The server application provides a web interface where all data can be viewed, listened to and audio recordings can be downloaded. The interface also allows to display statistical data, and it is possible to filter the data, e. g. by age, diagnosis, gender, etc. as well as to export some specially selected parameters to an Excel sheet.

To allow further data processing by artificial intelligence, various features were extracted from the audio files:

- Text transcription of speech recordings, with time marks for individual words and long and short pauses.
- Length of the whole recording and duration of the speech itself.
- Number of words, number of syllables.
- Speaking rate in words per minute, syllables per minute.
- Response time.
- Time needed for a correct answer in the naming test.
- Number and types of hesitations in speech.
- Features extracted using Neurospeech toolkit[89].
- Features extracted using OpenSmile toolkit[90] with Gemaps settings[91].
- Trill Embedding[92].

Annotation. To achieve that the recordings can be used for training specific acoustic and language models, collected audio files needed to be transcribed. First, automatic transcriptions were created using a speech recognizer developed at the Institute of Informatics of the Slovak Academy of Sciences. Subsequently, the automatic transcriptions were corrected by trained annotators using the program Transcriber 1.5.1[93] and adding labels for different acoustic events.

Hesitations are one of the most frequently observed parameters in spontaneous speech of patients with cognitive decline and/or dementia. However, the term hesitation is not used uniformly throughout various studies

AGE						
	AD	**AD-PD**	**HC**	**MCI**	**PD**	**SUM**
<50	—	—	66 (61)	—	13 (13)	**79 (74)**
50–59	—	—	341 (286)	—	22 (22)	**363 (308)**
60–69	6 (3)	1 (0)	459 (338)	6 (5)	62 (61)	**534 (407)**
70–79	35 (10)	1 (0)	320 (176)	32 (13)	54 (53)	**442 (252)**
>=80	45 (14)	—	135 (34)	23 (12)	24 (20)	**227 (80)**
missing data	1	—	2	1	—	**4**
SUM	**87**	**2**	**1323 (896)**	**62**	**175**	

Table 1. Distribution of EWA-DB according to age.

EDUCATION						
	AD	**AD-PD**	**HC**	**MCI**	**PD**	**SUM**
Primary	21 (2)	—	62 (20)	4 (1)	9 (8)	**96 (31)**
Secondary	45 (18)	2 (0)	707 (436)	26 (11)	108 (103)	**888 (568)**
University	20 (7)	—	552 (439)	31 (18)	58 (58)	**661 (522)**
missing data	1	—	2	1	—	**4**
SUM	**87**	**2**	**1323**	**62**	**175**	

Table 2. Distribution of EWA-DB according to education.

GENDER						
	AD	**AD-PD**	**HC**	**MCI**	**PD**	**SUM**
Female	63 (20)	1 (0)	939 (647)	40 (19)	71 (66)	**1114 (752)**
Male	23 (7)	1 (0)	382 (248)	21 (11)	104 (103)	**531 (369)**
missing data	1	—	2	1	—	**4**
SUM	**87**	**2**	**1323**	**62**	**175**	

Table 3. Distribution of EWA-DB according to gender.

focused on spontaneous speech analysis. While in some cases the term hesitation is used in a narrow sense, such as absence of speech during more than 30ms[94], in other cases it is used as an umbrella term[66] including various manifestations, as in silent pauses, filled pauses, but also verbal expressions such as false starts or restarts, interjections, or prolonged sounds of words. In our annotation convention, we were primarily guided by whether the hesitation was verbal or non-verbal. Non-verbal, but vocal expressions were tagged as [hez] - typically a schwa sound, but it could also be any other vocal expressions indicating word-finding difficulties. Silent pause tags were not introduced. Non-verbal sounds with a certain meaning, which did not indicate word-finding difficulties, were transcribed verbatim and tagged with a % symbol, e. g. %a: in the meaning of surprise or %hm in the meaning of thinking. Non-verbal sounds without any meaning were marked with a different tag [spk], e. g. in the case of mouth smacking or throat clearing. Of the verbal expressions, we paid attention to false starts and restarts, which were tagged by their joint tag + [ned], that was added to the beginning of the word being falsely started or restarted. Furthermore, we were paying attention to interjections, which were literally transcribed. We also considered sound prolongations to be a form of hesitation, as they can occur as results of word-finding difficulties. Such manifestations were given a separate tag – a colon was written after the sound concerned.

As hypokinetic dysarthria is present in many patients with PD, a reduced degree of speech intelligibility was expected. Words that were less intelligible were transcribed verbatim and placed in two parentheses ((word)). Words that were not intelligible at all were also placed in two parentheses and marked with as many "x" letters as the unintelligible word had syllables, e. g. ((x)) represents a monosyllabic unintelligible word. As changes in respiration are also documented in PD, the [ex] tag was used to indicate prolonged or loud exhalations.

We expected phonetic-phonological difficulties in both patients with PD[95] and AD[96]. Differences in pronunciation were observed and tagged using the forward slash symbol. First the target pronunciation was written down following the forward slash symbol and the actual pronunciation of the participant concerned. Thus, also slips of the tongue and phonemic or phonetic paraphasia were marked using this tag.

When working with the elderly and especially when collecting data from patients with neurodegenerative diseases, the input of investigators was necessary in some cases, e. g. to refine the instruction. Such entries were manually removed from the transcripts and the segment was marked as [inv].

The last tag included was [ruch]. This marked all background disturbances that could affect the correct capture of analyzed speech.

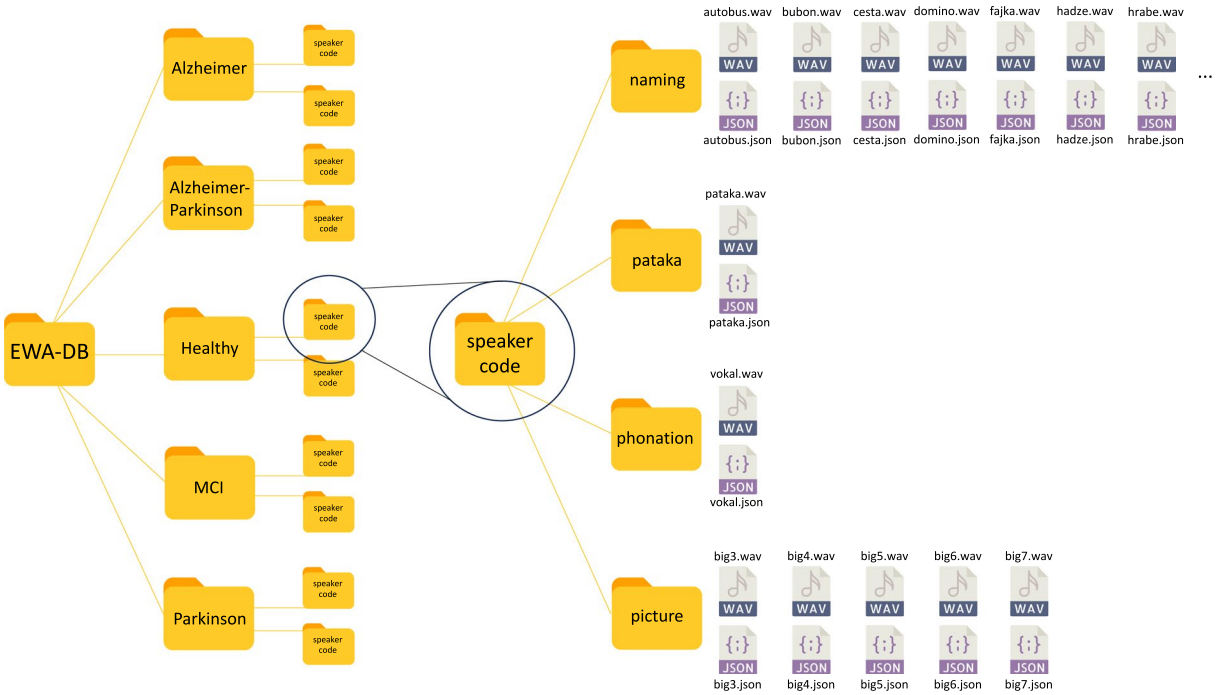

Fig. 3 Structure of EWA-DB.

Data Records

The EWA-DB speech database is publicly available at ELDA[97] under the name EWA-DB, and at ZENODO[98]. The database contains 1649 speakers, of which 87 are AD patients, 175 are PD patients, 2 speakers have a combination of AD and PD, 62 are MCI patients and 1323 are healthy controls (HC). However, when recruiting participants in clinical samples, or in the sample of healthy controls, the predefined inclusion criteria (see Inclusion and exclusion criteria) were followed. In several cases, neurologists or psychiatrists gave the diagnosis of AD, PD or MCI (or the diagnosis was refuted, and the participants were marked as healthy), but within the inclusion criteria the participants did not meet some conditions (e.g., MoCA scores). In such cases, the participants were not excluded and are still part of the database. If we were to consider precisely defined inclusion and exclusion criteria, the size of the database would be 1122 speakers. Within the database, for each speaker, there is information about meeting/not meeting the inclusion criteria. The distribution of the database in terms of age, education and gender are shown in Tables 1–3. We first present the total number of participants in each category, with the number of participants who met the inclusion criteria shown in parentheses.

Since not all participants agreed to the publication of their audio recordings, out of the total number of 1649 speakers, the database contains the recordings of 1003 speakers, i.e., those from whom we have written consent.

However, all 1649 speakers gave written consent to the processing of their recordings. The recordings of all participants were transcribed using automated speech recognition. Subsequently, the transcriptions were corrected and annotated by trained annotators (see Annotation). ASR transcription is available for every speaker (N = 1649), even for those speakers who did not give written consent to the publication of the audio recording. Manually annotated transcripts are available for 1502 speakers – for every speaker from the clinical sample and for the majority of speakers from the control sample.

Structure of EWA-DB. All files of the database are contained in a main folder named EWA-DB. The structure of this folder is depicted in Fig. 3.

After opening EWA-DB, additional folders will be displayed - one for each clinical group and one for the group of healthy controls. In each group of participants, there are subsequent folders for individual speakers, which are marked with randomly generated codes. After opening the folder of one speaker, folders for individual language tasks will be displayed - that is, separate folders for phonation, diadochokinesis (pataka), naming and picture description.

If the speaker gave consent the folders contain recordings as WAV files and additional JSON files (UTF-8 character encoding) containing ASR transcription, annotation, speaker information, Trill Embedding, acoustic features measured via Neurospeech and OpenSmile (Gemaps set). If the consent to publish recordings was not given, only a JSON file is provided.

The folder *phonation* contains the task of sustained vowel phonation.

The folder *pataka* contains the task for diadochokinesis.

The folder *naming* contains several files – one WAV file and one JSON file for every picture used in the object and action naming task. These files are marked with the correct naming response in the Slovak language. The name of each file with the corresponding English translation is provided as a supplementary material.

	AD		MCI		PD		HC	
Participants	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**
Primary	1	1	—	1	2	6	6	14
Secondary	4	14	3	8	59	44	115	322
University	2	5	8	10	42	16	127	312
Total participants	**7**	**20**	**11**	**19**	**103**	**66**	**248**	**648**
AGE average	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**
Primary	71.0	83.0	—	80.0	57.0	72.5	68.0	69.8
Secondary	73.8	81.1	72.7	77.1	64.7	70.3	61.6	69.6
University	78.5	81.4	75.6	76.3	68.7	63.4	62.8	60.8
Age total average	**74.7**	**81.3**	**74.8**	**76.8**	**66.2**	**68.8**	**62.4**	**65.4**
MOCA average	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**
Primary	19.0	20.0	—	25.0	28.0	26.3	28.0	27.1
Secondary	21.0	19.8	23.7	23.8	25.6	25.6	27.5	27.8
University	19.5	20.8	23.9	23.7	25.8	26.4	27.9	28.1
MOCA average	**20.3**	**20.1**	**23.8**	**23.8**	**25.7**	**25.9**	**27.8**	**27.9**

Table 4. Samples composition of the EWA database.

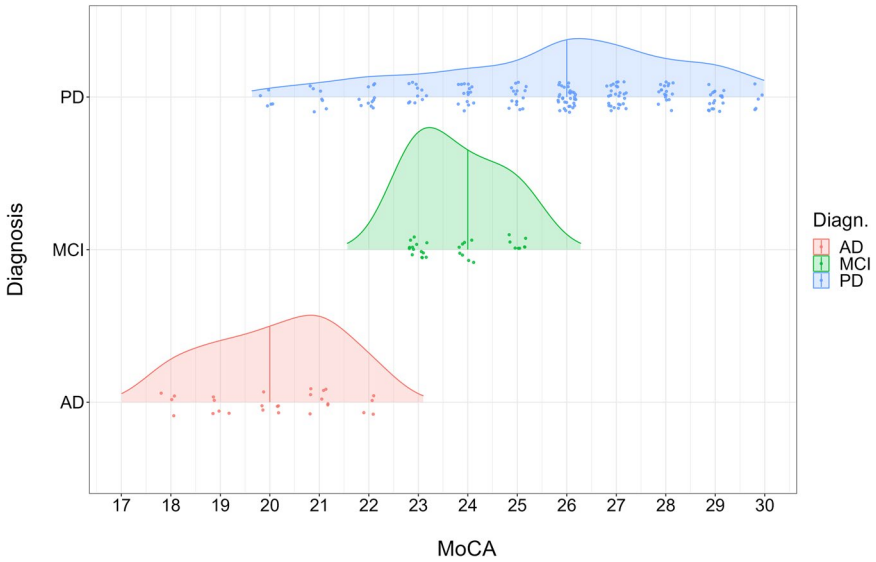

Fig. 4 Density distribution plots for clinical samples according to their MoCA score. The density distribution plot shows distribution in the Alzheimer's disease (AD), Parkinson's disease (PD) and mild cognitive impairment (MCI) clinical samples according to their cognitive screening MoCA scores.

The folder *picture description* contains the descriptions of three complex pictures named *big3*, *big 4* and *big 5* and two simple pictures named *big 6* and *big 7*.

Technical Validation

Characteristics of EWA-DB in terms of age, gender, education and MoCA score. The database contains a complete sample of 1122 participants meeting the inclusion criteria. In Table 4, for each group of patients with AD, MCI, PD and the group of healthy controls, the summary numbers of participants, the mean age and the mean MoCA score by education and gender are presented.

As shown in Table 4, the smallest number of participants is in the group of patients with AD (N = 27). The total number of MCI patients is (N = 30). There are fewer males than females in both the AD and MCI patient groups. The total number of patients with PD is (N = 169) and unlike the other groups, the number of men (N = 103) is higher than the number of women (N = 66). In our sample, the average age of patients with AD and MCI increases with higher education in men. On the contrary, in women, the average age decreases with education. In the case of primary and secondary education, the average age is higher for women than for men in all groups. There are no notable differences in mean MoCA scores within genders according to education.

The patient density graph for AD patients by MoCA score in Fig. 4 shows that the median of MoCA score is 20 and the largest number of AD patients have a MoCA score of 21 above the median. In MCI patients, on the contrary, the MoCA median is 24, but the largest number of patients have a MoCA score below 23. This suggests

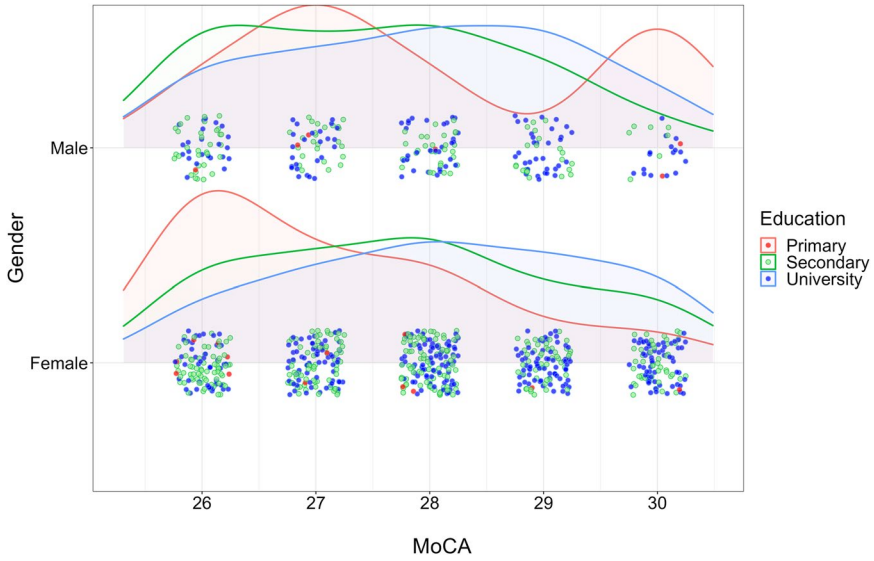

Fig. 5 Distribution curves for healthy control according to MoCA, gender, and education. The graph shows the distribution of healthy control participants according to three factors: cognitive screening MoCA score, gender (male or female), and education (primary, secondary or university).

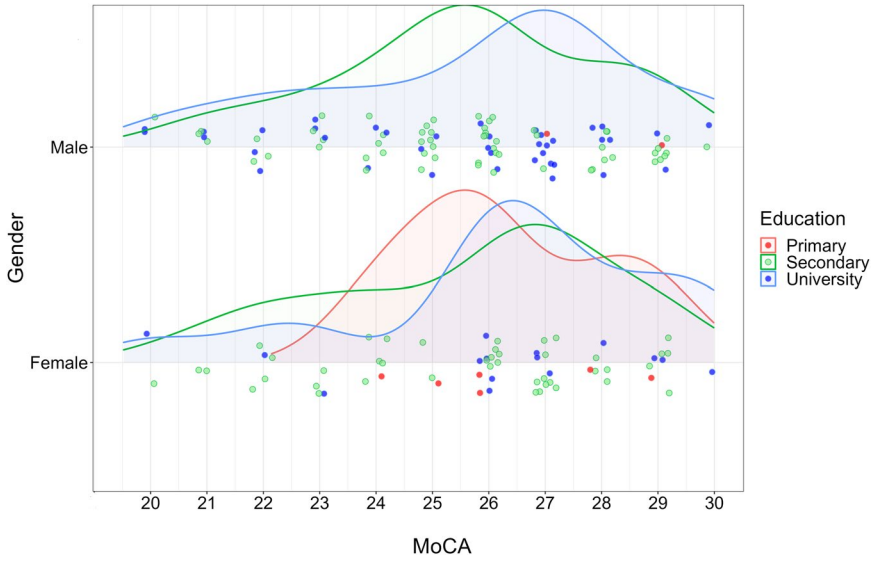

Fig. 6 Distribution curves for PD patients according to MoCA, gender, and education. The graph shows the distribution of the Parkinson's disease (PD) clinical sample according to three factors: cognitive screening MoCA score, gender (male or female), and education (primary, secondary or university).

that there is a natural transition of MCI patients to AD according to MoCA scores. The density distribution plot of PD patients in Fig. 5 shows that the median of MoCA score is 26 and the largest number of patients have a MoCA score of 26. The density distribution of PD patients is more elongated towards lower MoCA score values of 20.

Figure 5 shows the distribution of the healthy control group by MoCA score, gender, and education level, using the statistical software R. The coloured dots represent healthy participants according to their education (primary, secondary, and higher education), randomly distributed near their MoCA scores on both axes. A separate distribution is displayed for men and women by gender. Healthy participants are evenly distributed around a MoCA score of 28 for secondary and higher education. The proportion of participants with higher and secondary education is 49%, while the proportion of participants with primary education is 2%, with the predominant MoCA scores for primary education being 26 for women and 27 for men.

Figure 6 shows the distribution of PD patients by MoCA score, gender, and education level. The coloured dots represent PD patients according to their education (primary, secondary, and higher education), distributed near their MoCA scores. The proportion of men with PD is 61%, higher compared to women at 39%. Among

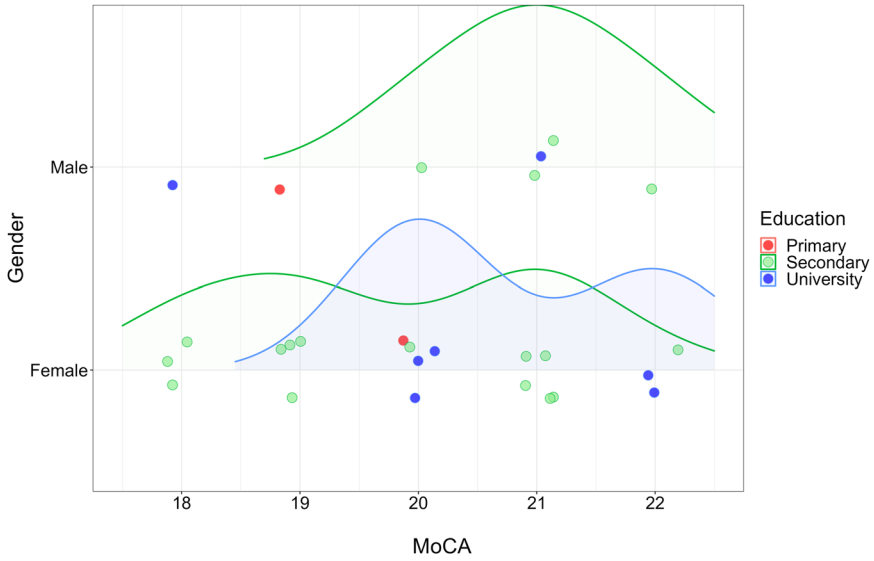

Fig. 7 Density graph of AD patient distribution according to MoCA, gender, and education. The graph shows the distribution of the Alzheimer's disease (AD) clinical sample according to three factors: cognitive screening MoCA score, gender (male or female), and education (primary, secondary or university).

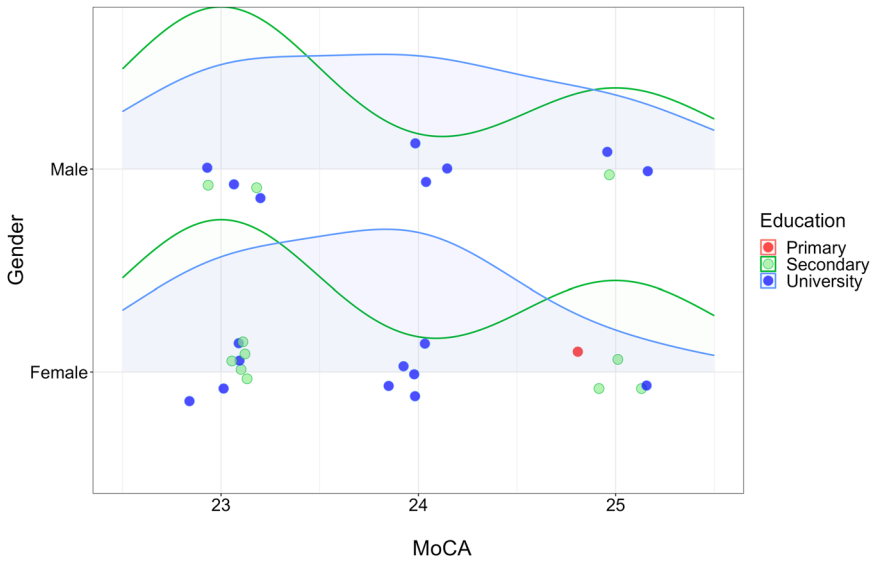

Fig. 8 Density graph of MCI patient distribution according to MoCA, gender, and education. The graph shows the distribution of the mild cognitive impairment (MCI) clinical sample according to three factors: cognitive screening MoCA score, gender (male or female), and education (primary, secondary or university).

women with PD, secondary education is significantly dominant at 67%. PD patients with higher education have a median MoCA score of 26.5, with lower MoCA scores ranging from 20 to 26. PD patients with secondary education are more evenly distributed around a median MoCA score of 26. The proportion of participants with primary education is 5%, mainly represented by women, while 61% have secondary education, and 34% have higher education.

The number of AD patients is small (N = 26), and the representation of patients can be directly observed in Fig. 7. There is a low representation of men with AD, with the MoCA median (21) being higher than the MoCA median in women (20). Most of AD patients have secondary education.

Similar to AD, the number of MCI patients is also small (N = 30), and the representation of patients can be directly observed in Fig. 8. The MoCA median (24) is the same for men and women, with the highest number of MCI patients with a university degree.

Quality control of recordings and transcripts. During the EWA project, recordings were made at various clinical sites. Especially at the beginning of the project, it occurred that during the recording there were

disturbing noises and sounds in the background, such as the siren of an ambulance, noise from the corridor, interruption by medical personnel, knocking on the door, or the administrator taking notes near the microphone during the testing procedure. Disturbing moments like these can have an impact on sound parameters and affect the process and results of machine learning. For this reason, a subjective evaluation of the acoustic quality of the recordings was performed. Every recording that was considered acoustically unsatisfactory was marked as low quality. Information about the quality of each recording is also provided in the JSON file.

The next phase of quality control took place in the transcription process. Automatic speech recognition was used to create transcriptions from speech recordings. These transcripts were consequently checked and corrected by trained annotators, most of whom were speech and language pathology students who already gained basic experience with creating transcriptions during their studies. Each annotator took part in a detailed training at the Institute of Informatics of Slovak Academy of Sciences. As part of this training, the annotators learned how to properly check, correct, and annotate transcripts and how to work with the Transcriber 1.5.1 program[93]. At first, the work of the annotators was checked by their supervisor, a speech and language pathologist, until the conclusion was made, that the annotators had adopted the established rules for annotation and could work independently. Annotators received transcripts as TRS files and corresponding recordings as WAV files. In addition to entering tags (see Annotation), the task of each annotator was to check the transcription and, in case of inconsistencies, to correct them. When annotating the picture description tasks, annotators divided the transcription into sentences, following intonation, semantics, and syntax. Out of the total number of 1649 transcripts, 1502 is manually annotated. If the annotation is available, it is included in the JSON file.

Dataset limitations. The dataset developed in this study presents several limitations that warrant consideration. Perhaps the most significant limitation is the sample size, particularly concerning the clinical sample. This is partly due to the absence of audio recordings for certain participants, a consequence of not obtaining informed consent for those recordings. On the other hand, all participants gave their consents to include transcripts of these recordings. It should also be noted that the metadata does not include the results related to the inclusion criteria, as these were not captured within the scope of the informed consent process. This omission could impact the interpretation of the data to some extent.

In addition to these limitations, the data collection process itself introduced a notable degree of variability. This variability is attributable to several factors. Firstly, the use of a large number of recording devices with differing technical specifications (various types of smartphones) likely contributed to inconsistencies in the quality of the recordings. Secondly, the recordings were made at different times of the day, which may have affected participant fatigue levels and, consequently, their performance. Moreover, the recordings were conducted in diverse environments, each with its own acoustic characteristics. These environmental differences likely impacted the acoustic quality of the recordings, leading to further variability in data quality.

While we have taken steps to mitigate these issues, it is important to acknowledge that these limitations could affect the generalizability and robustness of the study's findings. Future studies might consider standardizing recording conditions and obtaining comprehensive informed consent to minimize these limitations.

Code availability

No custom code has been used.

Received: 25 June 2024; Accepted: 26 November 2024;
Published online: 04 December 2024

References

1. Dubois, B. *et al.* Clinical diagnosis of Alzheimer's disease: recommendations of the International Working Group. *Lancet Neurol.* **20**, 484–496, https://doi.org/10.1016/S1474-4422(21)00066-1 (2021).
2. Postuma, R. B. *et al.* MDS clinical diagnostic criteria for Parkinson's disease. *Mov Disord.* **30**, 1591–1601, https://doi.org/10.1002/mds.26424 (2015).
3. Platero, C. & Toba, M. C. Alzheimer's Disease Neuroimaging Initiative. Predicting Alzheimer's conversion in mild cognitive impairment patients using longitudinal neuroimaging and clinical markers. *Brain Imaging Behav.* **15**, 1728–1738, https://doi.org/10.1007/s11682-020-00366-8 (2021).
4. Heinzel, S. *et al.* MDS Task Force on the Definition of Parkinson's Disease. Update of the MDS research criteria for prodromal Parkinson's disease. *Mov Disord.* **34**, 1464–1470, https://doi.org/10.1002/mds.27802 (2019).
5. World Health Organization. Dementia. https://www.who.int/news-room/fact-sheets/detail/dementia (2021).
6. Skorvanek, M., Feketeova, E., Kurtis, M. M., Rusz, J. & Sonka, K. Accuracy of rating scales and clinical measures for screening of rapid eye movement sleep behavior disorder and for predicting conversion to Parkinson's disease and other synucleinopathies. *Front Neurol.* **9**, https://doi.org/10.3389/fneur.2018.00376 (2018)
7. Boschi, V. *et al.* Connected speech in neurodegenerative language disorders: A review. *Front. Psychol.* **8**, https://doi.org/10.3389/fpsyg.2017.00269 (2017).
8. Martínez-Nicolás, I., Llorente, T. E., Martínez-Sánchez, F. & Meilán, J. Ten years of research on automatic voice and speech analysis of people with Alzheimer's disease and mild cognitive impairment: A systematic review article. *Front. Psychol.* **12**, https://doi.org/10.3389/fpsyg.2021.620251 (2021).
9. Goberman, A. M. & Coelho, C. Acoustic analysis of Parkinsonian speech I: Speech characteristics and L-Dopa therapy. *NeuroRehabilitation* **17**, 237–246 (2002).
10. Ash, S. *et al.* Impairments of speech fluency in Lewy body spectrum disorder. *Brain Lang.* **120**, 290–302, https://doi.org/10.1016/j.bandl.2011.09.004 (2012).
11. Rusz, J., Cmejla, R., Ruzickova, H. & Ruzicka, E. Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. *J. Acoust. Soc. Am.* **129**, 350–367, https://doi.org/10.1121/1.3514381 (2011).
12. Crescentini, C., Mondolo, F., Biasutti, E. & Shallice, T. Supervisory and routine processes in noun and verb generation in nondemented patients with Parkinson's disease. *Neuropsychologia* **46**, 434–447, https://doi.org/10.1016/j.neuropsychologia.2007.08.021 (2008).
13. Chen, L. *et al.* Improving the assessment of mild cognitive impairment in advanced age with a novel multi-feature automated speech and language analysis of verbal fluency. *Front. Psychol.* **11**, https://doi.org/10.3389/fpsyg.2020.00535 (2020).

14. Borson, S., Scanlan, J. M., Watanabe, J., Tu, S.-P. & Lessig, M. Improving identification of cognitive impairment in primary care. *Int. J. Geriatr. Psychiatry* **21**, 349–355, https://doi.org/10.1002/gps.1470 (2006).
15. Woodford, H. & George, J. Cognitive assessment in the elderly: A review of clinical methods. *QJM Int. J. Med.* **100**, 469–484, https://doi.org/10.1093/qjmed/hcm051 (2007).
16. Gullapalli, A. S. & Mittal, V. K. Early detection of Parkinson's disease through speech features and machine learning: A review. *ICT with Intelligent Applications* **248**, 203–212, https://doi.org/10.1007/978-981-16-4177-0_22 (2022).
17. Sevcik, A. & Rusko, M. A systematic review of Alzheimer's disease detection based on speech and natural language processing. *32nd International Conference Radioelektronika* **2022**, 01–05, https://doi.org/10.1109/RADIOELEKTRONIKA54537.2022.9764938 (2022).
18. Moro-Velazquez, L., Gomez-Garcia, J. A., Arias-Londoño, J. D., Dehak, N. & Godino-Llorente, J. I. Advances in Parkinson's disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects. *Biomed Signal Process Control* **66**, https://doi.org/10.1016/j.bspc.2021.102418 (2021).
19. Herd, C. P. *et al*. Comparison of speech and language therapy techniques for speech problems in Parkinson's disease. *Cochrane Database Syst Rev*. **2012**, https://doi.org/10.1002/14651858.CD002814.pub2 (2012).
20. Mueller, K. D., Hermann, B., Mecollari, J. & Turkstra, L. S. Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks. *J Clin Exp Neuropsychol* **40**, 917–939, https://doi.org/10.1080/13803395.2018.1446513 (2018).
21. Slegers, A., Filiou, R. P., Montembeault, M. & Brambati, S. M. Connected speech features from picture description in Alzheimer's disease: A systematic review. *J Alzheimers Dis* **65**, 519–542, https://doi.org/10.3233/JAD-170881 (2018).
22. Orozco-Arroyave, J. R, Arias-Londoño, J. D., Vargas-Bonilla, J. F., Gonzalez-Rátiva, M. C. & Nöth, E. New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. *International Conference on Language Resources and Evaluation*, (2014)
23. Arora, S., Baghai-Ravary, L. & Tsanas, A. Developing a large scale population screening tool for the assessment of Parkinson's disease using telephone-quality voice. *J Acoust Soc Am* **145**, https://doi.org/10.1121/1.5100272 (2019).
24. Kursun, O. *et al*. *Parkinson's Speech with Multiple Types of Sound Recordings*. UCI Machine Learning Repository. https://doi.org/10.24432/C5NC8M (2014).
25. Marek, K. *et al*. The Parkinson progression marker initiative (PPMI). *Prog Neurobiol* **95**, 629–635, https://doi.org/10.1016/j.pneurobio.2011.09.005 (2011).
26. Mendes-Laureano, J. *et al*. *Neurovoz: a Castilian Spanish corpus of parkinsonian speech*. arXiv preprint. https://arxiv.org/abs/2403.02371 (2024).
27. Jaeger, H., Trivedi, D. & Stadtschnitzer, M. Mobile Device Voice Recordings at King's College London (MDVR-KCL) from both early and advanced Parkinson's disease patients and healthy controls. *Zenodo*. https://doi.org/10.5281/zenodo.2867216 (2019).
28. Tsanas, A. & Little, M. *Parkinsons Telemonitoring. UCI Machine Learning Repository*. https://doi.org/10.24432/C5ZS3N
29. Becker, J. T., Boller, F., Lopez, O. L., Saxton, J. & McGonigle, K. L. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Arch Neurol* **51**, 585–594, https://doi.org/10.1001/archneur.1994.00540180063015 (1994).
30. Tripathi, T. & Kumar, R. Speech-based detection of multi-class Alzheimer's disease classification using machine learning. *Int J Data Sci Anal* **18**, 83–96, https://doi.org/10.1007/s41060-023-00475-9 (2024).
31. Luz, S., Haider, F., Fuente, S., Fromm, D. & MacWhinney, B. Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge. *Interspeech* https://doi.org/10.21437/Interspeech.2020-2571 (2020).
32. Pope, C. & Davis, B. H. Finding a balance: The Carolinas conversation collection. *Corpus Linguistics and Linguistic Theory* **7**, 143–161 (2011).
33. Liu, Z. *et al*. Dementia detection by analyzing spontaneous mandarin speech. *APSIPA ASC* https://doi.org/10.1109/APSIPAASC47483.2019.9023041 (2019).
34. Ambrosini, E. *et al*. Automatic speech analysis to early detect functional cognitive decline in elderly population. 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 212–216, https://doi.org/10.1109/EMBC.2019.8856768 (2019).
35. Fraser, K. C., Fors, K. L. & Kokkinakis, D. Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Computer Speech & Language* **53**, 121–139, https://doi.org/10.1016/j.csl.2018.07.005 (2019).
36. Thapa, S. *et al*. Detecting Alzheimer's disease by exploiting linguistic information from Nepali transcript. *International Conference on Neural Information Processing*, 176–184 (2020).
37. Lindsay, H., Tröger, J. & König, A. Language impairment in Alzheimer's disease – robust and explainable evidence for AD-related deterioration of spontaneous speech through multilingual machine learning. *Front aging neurosci* **13**, https://doi.org/10.3389/fnagi.2021.642033 (2021).
38. Rana, A. *et al*. Imperative role of machine learning algorithm for detection of Parkinson's disease: Review, challenges and recommendations. *Diagnostics* **12**, https://doi.org/10.3390/diagnostics12082003 (2022).
39. Pulido, M. L. B. *et al*. Alzheimer's disease and automatic speech analysis: A review. *Expert Sys Appl* **150**, https://doi.org/10.1016/j.eswa.2020.113213 (2020).
40. Rusz, J., Tykalova, T., Ramig, L. O. & Tripoliti, E. Guidelines for speech recording and acoustic analyses in dysarthrias of movement disorders. *Mov Disord.* **36**, 803–814, https://doi.org/10.1002/mds.28465 (2021).
41. Haspelmath, M., Dryer, M. S., Gil, D. & Comrie, B. *The World Atlas of Language Structures*. (OUP Oxford, 2005).
42. Tjaden, K. & Watling, E. Characteristics of diadochokinesis in multiple sclerosis and Parkinson's disease. *Folia Phoniatr Logop.* **55**, 241–259, https://doi.org/10.1159/000072155 (2003).
43. Duffy, J. R. *Motor Speech Disorders: Substrates, Differential Diagnosis and Management 4th ed*. (Mosby, 2020).
44. Gleichgerrcht, E., Fridriksson, J. & Bonilha, L. Neuroanatomical foundations of naming impairments across different neurologic conditions. *Neurology* **85**, 284–92, https://doi.org/10.1212/WNL.0000000000001765 (2015).
45. Li, D. *et al*. A color-picture version of Boston Naming Test outperformed the black-and-white version in discriminating amnestic mild cognitive impairment and mild Alzheimer's disease. *Front Neurol* **13**, https://doi.org/10.3389/fneur.2022.884460 (2022).
46. Faust, M. E., Balota, D. A. & Multhaup, K. S. Phonological blocking during picture naming in dementia of the Alzheimer type. *Neuropsychology* **18**, 526–536, https://doi.org/10.1037/0894-4105.18.3.526 (2004).
47. Rodríguez-Ferreiro, J., Menéndez, M., Ribacoba, R. & Cuetos, F. Action naming is impaired in Parkinson disease patients. *Neuropsychologia* **47**, 3271–3274, https://doi.org/10.1016/j.neuropsychologia.2009.07.007 (2009).
48. Kaplan, E. F., Goodglass, H. & Weintraub, S. *The Boston Naming Test. 1st Edition* (Lea & Febiger, 1978).
49. Kaplan, E. F., Goodglass, H. & Weintraub, S. *The Boston Naming Test. 2nd Edition* (Lea & Febiger, 1983).
50. Stenova, V. & Csefalvay, Z. *Faktory ovplyvnujuce lexikalne vyhladavanie v pomenovani obrazkov: Test pomenovania obrazkov*. [Factors Affecting Lexical Retrieval in Picture Naming: The Picture Naming Test]. (Slovenska asociacia logopedov, 2011).
51. Csefalvay, Z. & Valkovic, P. *Poruchy komunikacnych schopnosti pri Parkinsonovej chorobe*. [Communication Disorders in Parkinson's Disease]. (Comenius University, 2021).
52. Harry, A. & Crowe, S. F. Is the Boston Naming Test still fit for purpose? *Clin Neuropsychol.* **28**, 486–504, https://doi.org/10.1080/13854046.2014.892155 (2014).
53. Li, Y. *et al*. Culture effects on the Chinese version Boston Naming Test performance and the normative data in the native Chinese-speaking elders in Mainland China. *Front. Neurol.* **13**, https://doi.org/10.3389/fneur.2022.866261 (2022).

54. Viggiano, M. P., Vannucci, M. & Righi, S. A new standardized set of ecological pictures for experimental and clinical research on visual object processing. *Cortex* **40**, 491–509, https://doi.org/10.1016/S0010-9452(08)70142-4 (2004).
55. Bramão, I., Reis, A., Petersson, K. M. & Faísca, L. The role of color information on object recognition: A review and meta-analysis. *Acta Psychol (Amst)* **138**, 244–253, https://doi.org/10.1016/j.actpsy.2011.06.010 (2011).
56. Callahan, B. L. *et al*. Semantic memory impairment for biological and man-made objects in individuals with amnestic mild cognitive impairment or late-life depression. *J Geriatr Psychiatry Neurol.* **28**, 108–116, https://doi.org/10.1177/0891988714554708 (2015).
57. Duong, A., Whitehead, V., Hanratty, K. & Chertkow, H. The nature of lexico-semantic processing deficits in mild cognitive impairment. *Neuropsychologia* **44**, 1928–1935, https://doi.org/10.1016/j.neuropsychologia.2006.01.034 (2006).
58. Whatmough, C. *et al*. The semantic category effect increases with worsening anomia in Alzheimer's type dementia. *Brain Lang.* **84**, 134–147, https://doi.org/10.1016/S0093-934X(02)00524-2 (2003).
59. Taler, V., Voronkchikhina, A., Gorfine, G. & Lukasik, M. Knowledge of semantic features in mild cognitive impairment. *J Neurolinguistics* **38**, 56–70, https://doi.org/10.1016/j.jneuroling.2015.11.002 (2016).
60. Hwang, Y. K. *et al*. Diagnostic value of time-constrained naming test in mild cognitive impairment. *Dement Geriatr Cogn Disord* **44**, 171-181. https://doi.org/10.1159/000479149.
61. Vigliocco, G., Vinson, D. P., Druks, J., Barber, H. & Cappa, S. F. Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neurosci Biobehav Rev.* **35**, 407–26, https://doi.org/10.1016/j.neubiorev.2010.04.007 (2011).
62. Cotelli, M. *et al*. Action and object naming in Parkinson's disease without dementia. *Eur J Neurol.* **14**, 632–637, https://doi.org/10.1111/j.1468-1331.2007.01797.x (2007).
63. Péran, P. *et al*. Object naming and action-verb generation in Parkinson's disease: A fMRI study. *Cortex* **45**, 960–971, https://doi.org/10.1016/j.cortex.2009.02.019 (2009).
64. Fraser, K. C., Meltzer, J. A. & Rudzicz, F. Linguistic features identify Alzheimer's disease in narrative speech. *J Alzheimers Dis.* **49**, 407–422, https://doi.org/10.3233/JAD-150520 (2016).
65. Dijkstra, k, Bourgeois, M. S., Allen, R. S. & Burgio, L. D. Conversational coherence: Discourse analysis of older adults with and without dementia. *J Neurolinguistics* **17**, 263–283, https://doi.org/10.1016/S0911-6044(03)00048-4 (2004).
66. Szatlocki, G., Hoffmann, I., Vincze, V., Kalman, J. & Pakaski, M. Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. *Front Aging Neurosci*. **7**, https://doi.org/10.3389/fnagi.2015.00195 (2015)
67. Dushanova, J. *Diagnostics and Rehabilitation of Parkinson's Disease*. (Intechopen, 2011).
68. Ellis, C., Crosson, B., Gonzales Rothi, L. J., Okun, M. S. & Rosenbek, J. S. Narrative discourse cohesion in early stage Parkinson's disease. *J Parkinsons Dis* **5**, 403–411, https://doi.org/10.3233/JPD-140476 (2015).
69. Ellis, C., Fang, X. & Briley, P. Temporal aspects of global coherence during discourse production in early stage Parkinson's disease. *Advances in Parkinson's Disease* **5**, 41–49, https://doi.org/10.4236/apd.2016.53006 (2016).
70. Forbes-McKay, K. E. & Venneri, A. Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. *Neurol Sci.* **26**, 243–54, https://doi.org/10.1007/s10072-005-0467-9 (2005).
71. Liu, L. *et al*. Characteristics of language impairment in Parkinson's disease and its influencing factors. *Transl Neurodegener*. **4**, https://doi.org/10.1186/2047-9158-4-2 (2015).
72. Goodglass, H. & Kaplan, E. *The Assessment of Aphasia and Related Disorders*. (Lea & Febiger, 1972).
73. Cummings, L. Describing the Cookie Theft picture. Sources of breakdown in Alzheimer's Dementia. *Pragmat. Soc.* **10**, 153–176, https://doi.org/10.1075/ps.17011.cum (2019).
74. Wechsler, D. Wechsler Adult Intelligence Scale. Third Ed. *APA PsycTests* https://doi.org/10.1037/t49755-000 (1997).
75. Li, S. C. *et al*. Transformations in the couplings between intellectual abilities and constituent cognitive processes across the lifespan. *Psychol Sci.* **15**, 155–163, https://doi.org/10.1111/j.0956-7976.2004.01503003.x (2004).
76. Schroeder, D. H. & Salthouse, T. A. Age-related effects on cognition between 20 and 50 years of age. *Pers Individ Dif* **36**, 393–404, https://doi.org/10.1016/S0191-8869(03)00104-1 (2004).
77. Verhaeghen, P. Aging and vocabulary scores: A meta-analysis. *Psychol Aging* **18**, 332–339, https://doi.org/10.1037/0882-7974.18.2.332 (2003).
78. Bright, P. & van der Linde, I. Comparison of methods for estimating premorbid intelligence. *Neuropsychol Rehabil.* **30**, 1–14, https://doi.org/10.1080/09602011.2018.1445650 (2020).
79. Nasreddine, Z. S. *et al*. The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *J Am Geriatr Soc* **53**, 695–699, https://doi.org/10.1111/j.1532-5415.2005.53221.x (2005).
80. Bartos, A. & Raisova, M. *Testy a dotazniky pro vysetrovani kognitivnich funkci, nalady a sobestacnosti*. [Tests and Questionnares for the Assessment of Cognitive Functions, Mood and Self-sufficiency] (Mlada Fronta, 2015).
81. Mahoney, F. I. & Barthel, D. W. Barthel Index. *APA PsycTests* https://doi.org/10.1037/t02366-000 (1965).
82. Scogin, F., Rohen, N. & Bailey, E. Geriatric Depression Scale. In *Handbook of psychological assessment in primary care settings* (ed. Maruish, M. E.) 491–508 (Lawrence Erlbaum Associates Publishers, 2000).
83. Hajdúk, M. *et al*. *NEUROPSY: Štandardizácia neuropsychologickej testovej batérie na dospelej slovenskej populácii*. (Univerzita Komenského v Bratislave, 2021).
84. Spitzer, R. L., Kroenke, K., Williams, J. B. W. & Löwe, B. Generalized Anxiety Disorder 7 (GAD-7). *APA PsycTests* https://doi.org/10.1037/t02591-000 (2006).
85. Albert, M. S. *et al*. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* **7**, 70–79, https://doi.org/10.1016/j.jalz.2011.03.008 (2011).
86. Kulcsarova, K. *et al*. Comparison in detection of prodromal Parkinson's disease patients using original and updated MDS research criteria in two independent cohorts. *Parkinsonism Relat Disord* **87**, 48–55. https://doi.org/10.1016/j.parkreldis.2021.04.028.
87. https://dev.mysql.com/ MySQL.
88. High Performance Object Storage for Modern Data Lakes https://min.io/.
89. Orozco-Arroyave, J. R. *et al*. NeuroSpeech: An open-source software for Parkinson's speech analysis. *Digit Signal Process* **77**, 207–221 (2018).
90. Eyben, F., Wöllmer, M. & Schuller, B. Opensmile: the Munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM International Conference on Multimedia*, 1459–1462 (2010).
91. Eyben, F. *et al*. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing* **7**, 190–202 (2016).
92. Shor, J. *et al*. Towards learning a universal non-semantic representation of speech. Preprint at https://arxiv.org/abs/2002.12764 (2020).
93. Barras, C. *et al*. Transriber: development and use of a tool for assisting speech corpora production. *Speech Commun.* **33**, 5–22, https://doi.org/10.1016/S0167-6393(00)00067-4 (2001).
94. Hoffmann, I. *et al*. Temporal parameters of spontaneous speech in Alzheimer's disease. *Int J Speech Lang Pathol* **12**, 29–34, https://doi.org/10.3109/17549500903137256 (2010).
95. Elorriaga-Santiago, S. *et al*. Phonological processing in Parkinson's disease: a neuropsychological assessment. *Neuroreport* **24**, 852–855, https://doi.org/10.1097/WNR.0000000000000005 (2013).

96. Croot, K. *et al*. Phonological and articulatory impairment in Alzheimer's disease: a case series. *Brain Lang.* **75**, 277–309, https://doi.org/10.1006/brln.2000.2357 (2000).
97. Rusko, M. *et al*. *EWA-DB Early Warning of Alzheimer speech database*. https://catalog.elra.info/en-us/repository/browse/ELRA-S0489/ (2023).
98. *EWA-DB – Early Warning of Alzheimer speech database*. https://zenodo.org/records/10952480, https://doi.org/10.5281/zenodo.10952480 (2024).

Acknowledgements

This study was created in relation to the project EWA - Early Warning of Alzheimer (ITMS2014+: 313022V631), which was funded by the European Regional Development Fund and the project ALOIS - Diagnosis of Alzheimer's disease from speech using artificial intelligence and social robotics (APVV-21-0373), which was funded by the Slovak Research and Development Agency. The study was also funded by the Slovak Grant and Development Agency under contract APVV-22-0279 and by the EU Renewal and Resilience Plan "Large projects for excellent researchers" under grant No. 09I03-03-V03-00007.

Author contributions

Milan Rusko, Alfréd Zimmermann, and Eugen Ružický are the leaders of the research teams of the members of the consortium, they participated in the organization of the creation of the database and the collection of recordings. Matej Škorvánek and Petra Brandoburová were the scientific and clinical guarantors of the project, the main authors of the methodology, as well as the organizers and supervisors of the recording process. Richard Malaschitz and Marián Trnka provided the necessary technical solutions, designed the data collection and storage system, as well as data processing into the definitive form of the database. Róbert Sabo and Viktória Kevická were responsible for manual annotation of transcripts, namely recruiting, training, and supervising the work of annotators.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-04171-6.

Correspondence and requests for materials should be addressed to M.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.