General Psychiatry

Speech databases for mental disorders: A systematic review

Yiling Li,[1] Yi Lin,[1] Hongwei Ding,[1] Chunbo Li[2,3]

To cite: Li Y, Lin Y, Ding H, *et al*. Speech databases for mental disorders: A systematic review. *General Psychiatry* 2019;**32**:e100022. doi:10.1136/gpsych-2018-100022

Received 11 October 2018
Revised 20 March 2019
Accepted 31 March 2019

[1]School of Foreign Languages, Shanghai Jiao Tong University, Shanghai, China
[2]Shanghai Key Laboratory of Psychotic Disorders, Shanghai Mental Health Center, School of Medicine, Shanghai Jiao Tong University, Shanghai, China
[3]Center for Excellence in Brain Science and Intelligence Technology (CEBSIT), Chinese Academy of Sciences, Shanghai, China

Correspondence to
Professor Hongwei Ding; hwding@sjtu.edu.cn

ABSTRACT

Background The employment of clinical databases in the study of mental disorders is essential to the diagnosis and treatment of patients with mental illness. While text corpora obtain merely limited information of content, speech corpora capture tones, emotions, rhythms and many other signals beyond content. Hence, the design and development of speech corpora for patients with mental disorders is increasingly important.

Aim This review aims to extract the existing speech corpora for mental disorders from online databases and peer-reviewed journals in order to demonstrate both achievements and challenges in this area.

Methods The review first covers publications or resources worldwide, and then leads to the reports from China, followed by a comparison between Chinese and non-Chinese regions.

Results Most of the speech databases were recorded in Europe or the United States by audio or video. Some were even supplemented by brain images and Event-Related Potential (ERP) statistics. The corpora were mostly developed for patients with neurocognitive disorders like stutter and aphasia, and mental illness like dementia, while other types of mental illness such as bipolar disorder, anxiety, depression and autism were scarce in number in database development.

Strengths and limitations The results demonstrated that database development of neurocognitive disorders in China is much scarcer than that in some European countries, but the existing databases pave an instructive road for psychiatric problems. Also, the methods and applications of databases from the leading countries are inspiring for Chinese scholars, who are searching methods for developing a comprehensive resource for clinical studies.

BACKGROUND

The term 'mental disorder' is defined by the WHO that comprises a wide range of mental problems, with different symptoms, which are 'generally characterized by some combination of abnormal thoughts, emotions, behaviors and relationships with others. Examples are schizophrenia, depression, intellectual disabilities and disorders due to drug abuse.'[1] The prevention and treatment of mental disorders is essential for the promotion of mental health.[2]

Moreover, the term 'speech database' or 'corpus' is defined as a collection or body of knowledge or evidence; especially a 'collection of recorded utterances used as a basis for the descriptive analysis of a language'.[3] Speech databases can be applied into many linguistic and non-linguistic research fields, including discourse analyses, language acquisition, neuroscience, sociology, and psychopathology, and so forth. The databases of patients with mental disorders can offer statistical support for language research, such as the verbal productive and perceptive symptoms of stutterers, so as to provide suggestions for diagnosis and treatment.

The study of those patients' spoken languages is an essential approach to understand the mental activities of human brains, which can be learnt by artificial intelligence for early screening and diagnosis. However, we found that there are few speech databases in Asia compared with those in the western countries. Regarding the large number of people with mental disorders in Asia and the special characteristics of their languages, the significance of building speech databases in Asia is self-evident.

A study on global epidemiology of mental disorders revealed that there was a lack of data that report the prevalence of mental disorders and the mortality rate in low/middle-income countries (LMICs), which may be due to the lack of research support, funding and resources.[4] Similarly, a large-scale statistical collection on patients with mental disorders is mainly popular in Europe and the United States, while few databases or corpora of mental health are available in Asia.

In addition, many existing studies are merely tentatively inductive and comparative case studies in a single aspect, which falls behind in systematisation and standardisation. Since research based on corpora of mental disorders is gradually sprouting up, the speech database is transfering from unimodal to multimodal, which contains diverse records of data and helps people to understand individual behaviour with the help of effective and

cutting-edge analytical techniques, such as eye-tracking and event-related potential (ERP).

Therefore, from one respect, by combining corpora with mental illness, researchers can extract language features from a large number of common facts, and use linguistic, psychological, medical and other interdisciplinary means to reveal the expressive, behavioural and brain processing of the pathological groups. On the other hand, the mechanism of collection and extraction of language data for patients with mental disorders, to a certain extent, will establish a solid foundation for developing artificial intelligence regarding language diagnosis, rehabilitation and treatment in many regions, especially in LMICs.

Accordingly, the first aim of this article is to provide a review of the speech databases for psychotic language disabilities developed around the globe. The second aim is to extract and discuss the Chinese databases. In this way, it gives a comprehensive demonstration of the latest development of corpora for mental disorders worldwide.

METHODS

Data sources and search strategy

We searched five main databases including Web of Science, PubMed, Embase, PsycINFO and Cochrane Library with the English keywords *mental disorder*, *mental illness*, *corpus*, *database*, *language*, *data* and *speech*. We also searched three Chinese databases including China National Knowledge Infrastructure (CNKI), Wanfang Data and the VIP Database with the corresponding Chinese keywords *mental disorder, mental disease, corpus, database, language disability* and *articulatory disability*. Furthermore, other online resources that provide conference papers, registered trials and ongoing projects related to the topic were also taken into consideration. Moreover, tests or trials on neurological disorders were also inspected for a wider range of reference. Those resources include TalkBank,[5] a professional online corpus initiated by Carnegie Mellon University; the International Standard Randomised Controlled Trial Number (ISRCTN) registry,[6] a primary clinical trial registry recognised by the WHO and International Committee of Medical Journal Editors that accepts all clinical research studies; the International Conference on Language Resources and Evaluation (LREC),[7] an international conference organised by the European Language Resources Association; and WorldWide Science,[8] a science website comprised of different scientific databases and portals around the globe. Table 1 gives an example of the search terms in Cochrane Library.

Table 1 Search terms in Cochrane Library

	Search range	Keywords
	Title, Abstract, Keyword	Mental disorder
AND	Title, Abstract, Keyword	Speech
AND	All text	Database

We attempted to review papers and databases that are relatively novel and no older than 15 years. Hence, the range of dates of the research studies selected was between 1 June 2003 and 1 June 2018. The final search was run on 20 December 2018. As shown in figure 1, a total of 3310 English studies from five main English databases (Web of Science, Embase, PubMed, PsycINFO and Cochrane Library) and four other supporting databases (TalkBank, ISRCTN, LREC, WorldWide Science), as well as a total of 58 studies from three main Chinese databases (CNKI, Wanfang and VIP) were obtained.

Inclusion and exclusion criteria

The selection criteria qualify studies included in the review from the global perspective and Chinese perspective, respectively. As for the global research background, the studies were considered eligible if they met the following criteria: (1) including patients with neurological disorders; (2) recording speeches by audio or video; and (3) functioning as public research resources for other researchers. Conversely, they were considered unqualified if they had the following characteristics: (1) targeting on ordinary people without neurological diseases; (2) transcribing speeches into written words rather than audio records; and (3) functioning as private supporting materials inaccessible to public usage. As for the Chinese research, databases from other countries and regions were excluded.

Study screening and data extraction

The literature screening process of this review is shown in figure 1. The search and screen were performed independently by two authors (YLL, YL) using the search terms above in the online databases. The first step was to eliminate repeated studies. The second step was to screen the literature by checking the titles and abstracts of the studies according to the selection criteria mentioned above. The third step was to read the full text of the remaining literature and to further remove the literature according to the exclusion criteria. If the two search results were different, the two researchers reviewed the literature together and analysed the reasons for the differences. If the opinions were still conflicting, a third person (HWD) would examine and make the final decision. The database table was developed by the two researchers (YLL, YL) for description of the selected speech databases. The listed contents including the name of the speech database, establishment year, language, average age, number of subjects, location, media type, description and number of citations were shown in table 2. Among them, the number of citations was captured from Google Scholar for its broad coverage of journals and publications on 20 December 2018.

Quality assessment of included studies

For the quality assessment of included speech databases, we adopted the requirements for the establishment of the Corpus of Dutch Aphasic Speech (CoDAS),

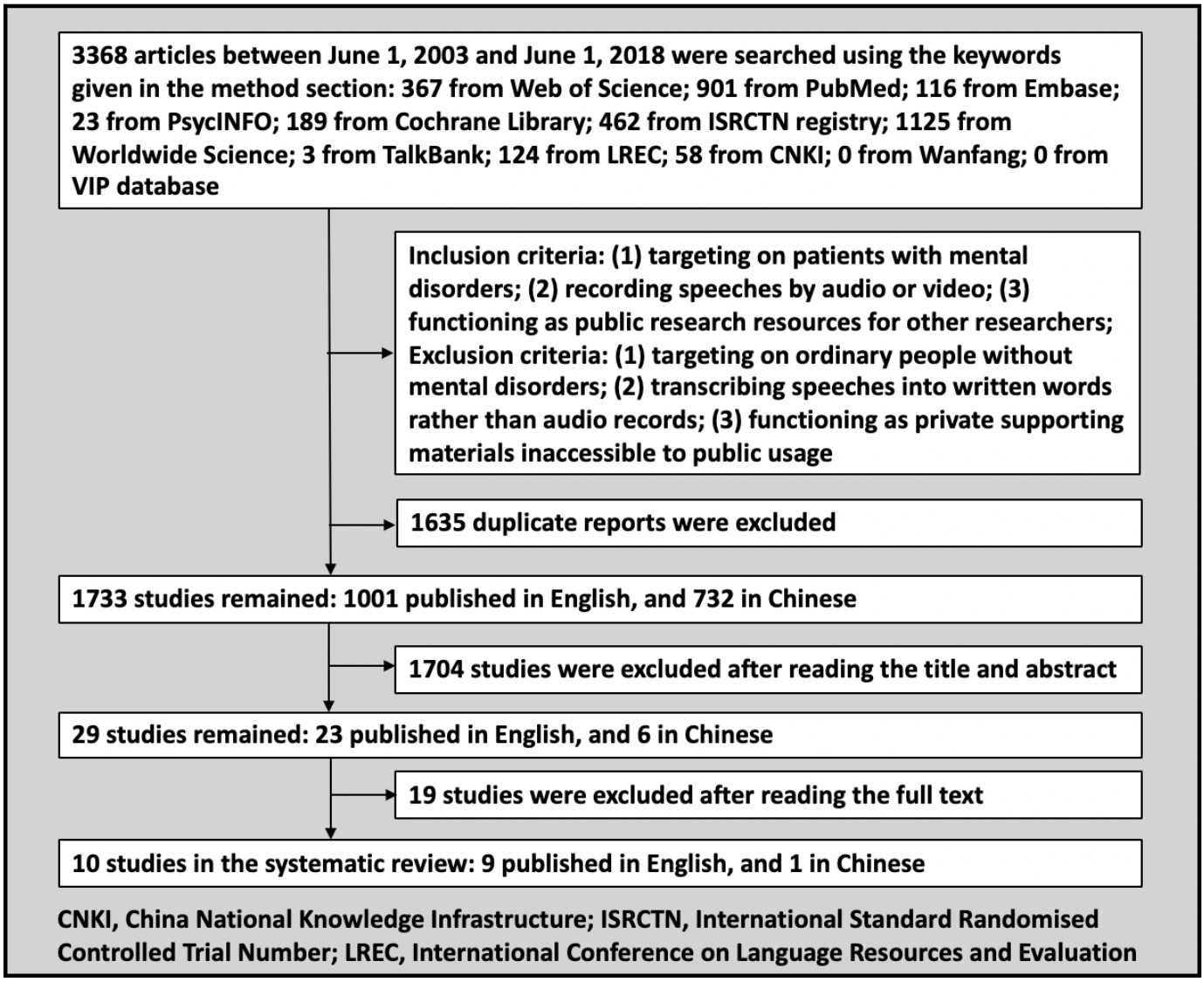

Figure 1 Flowchart of literature search and screening.

for it sets standards of text types, metadata and annotation that aphasia speech corpora should fulfil, and can be extended to a wider research field for other mental disorders. Accordingly, a corpus on language and speech processing of patients should fulfil at least the following requirements: (1) it should constitute a reasonable sample of patients within the region; (2) the speech recordings should be well documented with metadata about the subjects; and (3) the corpus should also include linguistic information such as part-of-speech tags, syntactic and prosodic annotation, as well as phonetic transcription.

The included speech databases went through the first step of basic requirements assessment accomplished by two investigators (YLL, YL). For the second step, the remaining databases were ranked according to the number of subjects, speech length, media type and number of citations, so that it could offer a reference for other researchers. The order of speech databases in table 2 was ranked by the number of citations. In this process, two other investigators (HWD, CBL) would be involved to discuss and resolve if there were any inconsistent opinions between the first two investigators.

RESULTS

Characteristics of included studies

First, we investigated the studies from the global perspective, which combines those from China and those from the other countries for the following discussions. Figure 1 demonstrates the procedure of literature selection.

First, we found 3368 articles between 1 June 2003 and 1 June 2018 (367 from Web of Science, 901 from PubMed, 116 from Embase, 23 from PsycINFO, 189 from Cochrane Library, 462 from ISRCTN registry, 1125 from World-Wide Science, 3 from TalkBank, 124 from LREC and 58 from CNKI). No result was found in Wanfang Data or VIP Database. Among them, we selected eligible studies by analysing the studies' titles and abstracts of the total

Table 2 Description of nine speech databases and their subsidiaries

Rank	Corpus	Year	Language	Average age	Subjects, n	Speech length (hour)	Location	Media type	Description	Citations, n
1	AphasiaBank					12				170
	Cantonese	2016	Cantonese	–	7/2		Hong Kong, China	Video	Native Cantonese speakers with stroke-induced aphasia	
	Croatian	2017	Croatian	–	10/10		Zagreb, Croatia	Video	Native Croatian speakers with stroke-induced aphasia	
	French	2016	French	–	11/14		France	Audio	Native French speakers with aphasia	
	Italian	2011	Italian	–	10		USA	Video	Native Italian speakers with aphasia	
	Mandarin	2015	Mandarin	45	9		China	Video	All patients with Mandarin as L1 and the aetiology is cerebral vascular accident (CVA)	
	Spanish	2011	Spanish	–	4		USA	Video	Communication impairments by monolingual and bilingual speakers of Spanish and/or English	
2	WRAP		English	54	64/200	30–40	USA	Audio	Connected speech problems of patients with dementia	168
3	Orozco-Arroyave Database	2014	Spanish	62; 60	50/50	>150	Spain	Audio	Speech recordings of patients with Parkinson's disease and healthy controls	57
4	DementiaBank					70–80				17
	English Holland	2016	English	68; 72	2		USA	Video	Individuals with Alzheimer's disease—language tasks from a Telerounds presentation	
	English Kempler	2016	English	81	6		USA	Audio	Individuals with Alzheimer's disease—conversation and Cookie-Theft picture descriptions	
	English Pitt	2016	English	–	208/104		USA	Audio	Dementia and control data for four language tasks from a large longitudinal study	
	English PPA DePaul	2016	English	66	1		USA	Video	Individual with primary progressive aphasia longitudinal data	
	English PPA Hopkins	2016	English	–	36		USA	Audio	Individuals with primary progressive aphasia data	
	German PPA	2016	English	–	–		Germany	Audio	Primary progressive aphasia data	
	Mandarin_Lu	2016	Mandarin	–	52		China	Audio	Individuals with dementia data	
	Spanish PerLA	2012	Spanish	–	21		Spain	Audio	Individuals with Alzheimer's disease and dementia data	
	Taiwanese_Lu	2016	Taiwanese	–	16		China	Audio	Individuals with dementia	

Continued

Table 2 Continued

Rank	Corpus	Year	Language	Average age	Subjects, n	Speech length (hour)	Location	Media type	Description	Citations, n
5	Cambridge Cookie-Theft Corpus	2010	English	54	87/227	41.5	Cambridge	Audio, brain scans	Individuals who have suffered from brain injury given the language task of picture description	9
6	CoDAS	2006	Dutch	54	6	0.5	Netherlands and Flanders	Audio	A pilot study of six aphasic speakers with two levels of annotation: an orthographic-phonetic transcription and a part-of-speech (POS) tagging	2
7	GREECAD	2016	Greek	55	72/28	1	Athens	Audio	An annotated Greek Corpus of Aphasic Discourse	1
8	FluencyBank					46				1
	POLER	2013	English	7	25/25		Washington, DC	Audio	Children with epilepsy and controls	
	IISRP	2005	English	4	100/50		USA	Audio	Seminal study of children who stutter with controls	
	Ratner	2012	English	3	23/15		USA	Audio	Children who stutter and controls	
	Voices	2006	English	42	12		USA	Video	Interviews from the Voices of Stuttering project	
	Ulm	1997	German	6	94		Ulm	Audio	Children who stutter from Ulm	
9	DAIC-WOZ	2014	English	–	–	50	USA	Video	Anxiety, depression and post-traumatic stress disorder in University of Southern California	124

Year denotes the establishment of the speech database.
n refers to the number of subjects and citations.
– denotes information not available or the project is still ongoing with an increasing number of subjects.
/ denotes the number of patients proportioned to the number of controls.
CoDAS, Corpus of Dutch Aphasic Speech; DAIC-WOZ, Distress Analysis Interview Corpus-Wizard of Oz; GREECAD, Greek Corpus of Aphasic Discourse; POLER, Plasticity of Language in Epilepsy Research; WRAP, Wisconsin Registry for Alzheimer's Prevention.

results and 1635 duplicate reports were excluded, with 1733 studies remaining in which 1001 were published in English and 732 in Chinese.

Second, 1704 studies were eliminated after reading the title and abstract. The reasons may be due to the misinterpretations of 'corpus' as 'an organ' or 'corpus' as 'a collection of writings', or due to the lack of audio or video records in the corpus and other irrelevant topics.

Third, we selected the studies by reading the full text to check if the studies focus on case studies with a corpus of ordinary, or non-targeted participants; or the studies of programming, annotation and analysis of existing corpora. Moreover, studies with statistics only and exclusively for language tests, education levels and literacy were also removed. For instance, Yu and colleagues[9] conducted a unique correlate analysis of patients with schizophrenia in their disorganized speech, with the Scale for the Assessment of Thought, Language and Communication, in which Word Fluency Test was used without acquiring speech samples. They found that patients with schizophrenia got much higher scores on the poverty of content of speech, distractible speech, tangentiality, derailment, and so forth. Of those studies, raw materials of speech or their transcriptions were not provided or systemised for further research. During this stage, 19 studies were excluded.

Finally, 10 unduplicated studies of speech databases were included in the analysis, nine in English and one in Chinese. All of them met the criteria: (1) targeting on patients with neurological disorders; (2) recording speeches by audio or video; and (3) functioning as public research resources for other researchers. All the corpora are accessible with or without charge dependent on corresponding contributors. The corpora and their subsidiary databases mainly apply the study methods of longitudinal tracking, language tasks or clinical analysis with a range of languages from English to German, Spanish, Croatian, Japanese, Mandarin, and so forth.

Comparison of the global and Chinese studies

Speech databases from the globe

The prominent contributor is TalkBank, which consists of the AphasiaBank, the DementiaBank, and the FluencyBank as shown in table 2. It is a multilingual corpus with lists of different research categories for resource sharing. Other contributions are leading national or international programmes on patients with mental disorders whose speech transcriptions, audio or video recordings and brain recordings have been organised into publicly available speech databases.

TalkBank

Established in 2002, TalkBank is a multilingual corpus managed and maintained by Brian MacWhinney, a professor of Psychology and Modern Languages at Carnegie Mellon University. Its goal is to provide resources for fundamental research in the field of human communication. In the corpus, DementiaBank is a shared database of communicative characteristics of patients with dementia. For instance, Becker and colleagues[10] have established Pitt Corpus in which written transcripts and audio files were were collected for administrating a protocol by the Alzheimer and dementia study at the School of Medicine, University of Pittsburgh. Similarly, AphasiaBank is an open database of multimedia approaches for the study of communication among patients with aphasia. For this database, a larger variety of languages, such as Croatian (Kraljević and colleagues[11], 2017) and French (Colin and Le Meur[12], 2017) are included. Moreover, FluencyBank focuses on children or adults with communicative problems, such as stutter, late talking and disfluency. In this speech database, Plasticity of Language in Epilepsy Research was a project conducted by Gaillard *et al*,[13] who selected 25 children with epilepsy compared to unaffected peers with the same age and gender in Children's National Medical Center in Washington, DC. Transcripts included the narrative task, which asked each child to give a story to the book *Frog, where are you?* Yairi and Ambrose[14] investigated 88 young children for the onset of stuttering. This includes tests of speech, language, hearing, motor skills, intellectual functioning and emotional reactions, as well as audio and video recordings, thorough case histories and familial pedigrees, which were conducted regularly for more than 12 years. Those two studies were carried out in English in the United States. Though aphasia and stutter do not necessarily imply a mental illness or impairment in intelligence, they might be attributed to neurogenic damage, such as a stroke or a brain tumor, but they are highly related to psychiatric diseases, as both of them are manifestation of neuro disruptions. To some extent, the investigation on databases of neurogenic illness may well provide valuable englightenment for mental disorders.

Independent speech corpora

As for independent aphasic speech corpora, Westerhout and Monachesi[15] developed a corpus containing aphasic speech recordings - the CoDAS with six participants. In this pilot study, they introduced basic requirements in terms of text types, metadata and annotation levels that this corpus should fulfil. Furthermore, they investigated the challenges that aphasic speech raises in orthographic transcription and part-of-speech tagging, thus providing an important direction for corpus developers.

The Cambridge Cookie-Theft Corpus is an audio corpus consisting of transcripts of speech by two tasks: the Boston Cookie-Theft picture description task[16] and a spontaneous speech task, which contains a semiprompted monologue (where the participants were asked to answer general non-intrusive questions about their daily lives and hobbies), and/or free speech (where an initial question was asked with no requirement of secondary prompting). Interviews on patients aged 19–89 years were matched with a comparable number of healthy individuals (aged 20–89 years). What distinguishes it from other corpora is that it also contains structural brain images of 32% of

the patients and 18% of healthy individuals. Research in language impairment is highly supported by this type of distinctive data that helps to facilitate analysis of natural language processing and corpus linguistics techniques.[16]

It was followed by an annotated Greek Corpus of Aphasic Discourse[17] with 72 patients and 28 controls aged 39–71 years under a large multidisciplinary project 'THALES-Levels of impairment in Greek aphasia: relationship with processing deficits, brain region, and therapeutic implications' This project aims to collect, annotate, document, and analyze the spoken discourse of Greek speakers with aphasia.

In addition, other corpora like the Orozco-Arroyave Database[18] provide speech recordings of 50 patients with Parkinson's disease compared by their healthy controls. The Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ) database is part of a larger corpus, the DAIC, which contains both audio and video recordings of patients with anxiety, depression and post-traumatic stress disorder (PTSD).[19] Finally, the Wisconsin Registry for Alzheimer's Prevention (WRAP) [20]is a longitudinal study containing recordings and transcriptions of 264 participants (200 cognitively healthy and 64 with early mild cognitive impairment [MCI]). It was developed for determining if participants with very early, subclinical memory declines were connected with declines in their language.

Overall, all of the nine databases were recorded by audio; four were also available in the form of video: DementiaBank, FluencyBank, AphasiaBank and DAIC-WOZ. One had additional brain scans: the Cambridge Cookie-Theft Corpus.

Filtered studies

The majority of the filtered studies failed to meet the criteria of public accessibility. Yet some of them offered inspiring possibilities for future directions of developing speech databases.

First, various test models and media types were applied. For example, Conroy[21] carried out cognitive and mental tests of 50 people with relapsing-remitting multiple sclerosis to find out their difficulties in retrieving words. The cognitive models applied the Addenbrooke's Cognitive Examination-Revised, the picture naming task (International Picture Naming Project, IPNP), the National Adult Reading Test (IPNP), and so on, accompanied by a one-time MRI scan and an in-depth neuropsychological assessment. In addition, Chakraborty and colleagues[22] investigated into patients with schizophrenia who were unable to identify emotion from voice due to their incapability in understanding low-level acoustic features, such as pitch, intensity, frequency, and so forth. The Negative Symptom Assessment-16 scale[23] rating method evaluated the behavior of those 78 participants (52 patients and 26 controls). Patel and colleagues[24] compared the speaking abilities of patients with amyotrophic lateral sclerosis with the healthy controls. They applied three speech categories (including task-rehearsed speech, spontaneous speech and repeated word) for eight patients and eight controls. Their speech was recorded by audio accompanied by an articulatory movement data recorder—electromagnetic articulography.

Second, self-evaluation was added to data collection. Volkmer[25] tried to find out whether the communication training program of Better Conversations with Primary Progressive Aphasia improves communication strategies, self-efficacy, and quality of life. In this case, speech and language therapists asked the patients to analyze the video samples of their own conversation to identify points that facilitate or hinder communication. Similarly, Beitchman and colleagues[26] launched a 26-year cohort study following up children with language problems from ages 5 to 31 years old. Dimensional psychosocial self-report measures were also adopted to determine whether the mental evaluation outcomes of 31 young patients are different from the outcomes of normal people.

Third, Dudy and colleagues[27] designed a computer-assisted pronunciation training program that is potentially to be a highly effective teaching aid for phonologically disordered preschool and school-age children. Effective pronunciation training requires prolonged supervised practice and interaction. In this case, 90 children aged 4–7 years (μ=5.3, σ=1.3) were recruited for the design of a speech corpus containing 53 simple words (eg, 'house', 'tree', 'window') elicited from describing images with the assistance of the speech-language pathologist.

Above all, considering the characteristics of practical and economic efficiency of clinical trials in medication, most of them have simplified and digitalised data collection for their own goals, regardless of the development of a sharable systematic database accessible for following researchers. Moreover, they also provide illuminating approaches for speech database development with the help of test models, neurotechnology and computer-aided technology.

Employment of the speech databases

The cookie-theft stimuli, the single black and white picture of the *Cookie-Theft* from the Boston Diagnostic Aphasia Examination Battery,[28] is one of the most used stimuli for the assessment of language production, and has been extended in many developmental projects of speech databases in addition to the Cambridge Cookie-Theft Corpus. Other picture description databases, referring to the *cookie-theft* stimuli, have also been introduced for broader speech samples. Boschi and colleagues[29] gave a review on the tasks of picture description, story narration and interview, for finding out possible different contributions to the assessment of different linguistic domains. Images such as the *picnic* picture[30] were used to develop a system that has 86.1% accuracy to predict early signs of cognitive decline; and the Picnic scene of Western Aphasia Battery[31] or the tales of *Cinderella*, *Snow White and the Seven Dwarfs* and the *Little Red Riding Hood*[32 33] have also been used in various studies but to a much lesser extent.

As for the DementiaBank, various studies in the past have employed this database. For instance, Fraser and

colleagues[34] achieved over 81% differentiation accuracy in distinguishing between participants with Alzheimer's disease (AD) and healthy controls by analyzing linguistic variables from the written scripts and acoustic variables from related audio files; Orimaye and colleagues[35] applied DementiaBank to learn lexical, syntactic, and *n-gram* linguistic biomarkers to distinguish the people of potential AD from the healthy controls. Their best diagnostic model could significantly distinguish the two groups using Support Vector Machines; Masrani and colleagues[36] used the DementiaBank data to prove that the AUGMENT domain adaptation algorithm, which improved the F-measure by more than 7% over models trained on MCI data, is better than all current baseline algorithms.

As for the AphasiaBank, Boyle[37] applied the AphasiaBank stimuli to examine the test–retest reliability for word-retrieval errors in narration by individuals with aphasia. The results showed that the test method was unreliable under each narrative subgenre. Furthermore, some patients were evaluated to find out whether the observed profiles were varied by race-ethnicity. Ellis and Peach[38] conducted a retrospective study on a convenience sample of persons with aphasia (PWA) extracted from AphasiaBank. In multivariate comparisons under certain age and education criteria, black PWAs exhibited lower scores in word fluency (5.5 vs 7.6; p=0.015), auditory word recognition (49.3 vs 53.3; p=0.02) and comprehension of sequential commands (43.7 vs 53.2; p=0.017) when compared with white PWAs.

Based on WRAP, Mueller and colleagues[39] showed that participants with early MCI status had a faster decline in speech fluency and semantic cognition than those normal peers. Dham *et al*[40] presented their multimodal feature extraction and decision-level fusion approach for detecting depression automatically. Features were extracted from the provided DAIC-WOZ database. The model proved to cross the provided baseline on validation data set by 17% on audio features and 24.5% on video features. Similarly, Cummins and colleagues [41] presented key factors that may reveal gender differences in the significance of depression on vowels' formant features through the application of DAIC-WOZ.

Speech databases from China

Since speech databases from China are still under development or yet to be developed, the earliest database development in Asia was spotted in Russia, Pakistan, India, Israel, etc.

Specifically, Rahman and colleagues[43] developed an interventional method called 'PASS' (Parent-mediated Intervention for Autism Spectrum Disorders in South Asia) by applying videos of parents and children's interaction at play. They studied 119 children with potential autism spectrum disorder (ASD) to come up with a comprehensive strategy to improve early detection. Bersudsky and colleagues[44] tested on 16 Russian immigrants to Israel aged 33–53 years, eight with schizophrenics and eight healthy immigrants for revealing differences between the sick and healthy peers. Short interviews around 10 min each were conducted, transcribed and coded for syntactic, lexical and pragmatic measures. Yu and colleagues[45] investigated into factors that led to the vulnerability of neuropsychological function in patients with Parkinson's disease at the early stage. Accordingly, tests of memory, attention, visuospatial, and language functions, etc. were adopted.

As for China's studies, five were left for further consideration. Those studies met the qualifications of (1) containing over 10 participants of patients with mental disorders in total, and (2) possessing original speech information by audio or video. Yet like the studies excluded from the non-Asian perspective, they applied test models, parallel data and small-scale video or radio recordings for private usage.

Sah and Torng[46] investigated the ability of Mandarin-speaking children with ASD who used mental state terms in narratives. The story-telling data were from 16 children with ASD and 16 normal children using the story *Frog, where are you?* collected by audio and video. Wu and colleagues[47] implemented a multimodal corpus research on patients with AD through on-site ad hoc recording and video recording. The comparative statistics were recorded in dialogue exchanges of normal old people of the same age and gender. They employed audio and video processing tools such as PRAAT[48] and EUDICO Linguistic Annotator (ELAN)[49] to perform simultaneous multilevel segmentation and labelling of the original materials. TalkBank was included in the result for its component of Asian languages. One of its contributors, the Cantonese AphasiaBank Database,[42] is currently a large cooperative project between Hong Kong and America. It provides abundant data of Cantonese speakers with aphasia and healthy controls in behavior, which includes years of linguistic, gestural and prosodic data collection and analyses of healthy Cantonese speakers as well as subjects with language disfunction secondary to left hemisphere stroke. The database has been and will continue to be periodically updated, aiming to help investigate how language abilities are affected by a neurological event like stroke or be applied to train speech-language pathologists or other related professionals.

434445

46474849

Employment of the speech databases

Most of those databases above have not been applied for experimental research. However, the Cantonese AphasiaBank Database is gradually being adopted among Hong Kong researchers. For instance, Fung and colleagues[50] studied the main concept analysis (MCA),[51] a content-based analytic approach that focuses on the quantification of presence, accuracy and completeness of oral discourse by PWAs to add MCA analysis to Cantonese AphasiaBank, and to examine the effects of age, gender, educational level and genre type on

discourse performance in unimpaired speakers, and so forth. Language samples of 105 PWAs and 150 unimpaired native Cantonese speakers were extracted from Cantonese AphasiaBank.

Lee and Kong[52] presented an investigation on the automatic speech recognition (ASR) system, a method of automatic speech assessment for Cantonese-speaking aphasic patients, which is a leading approach to extract robust text features based on word embedding methods. Speech recordings from 101 unimpaired speakers (about 12.6 hours) in the Cantonese AphasiaBank database were extracted to train the language models and acoustic models of the ASR system.

DISCUSSION

Main findings

The speech databases mentioned above are significant to different extents based on their rankings. In terms of the number of subjects, the Cambridge Cookie-Theft Corpus ranked first with 87 patients and 227 controls, followed by English Pitt that contains 208 patients and 104 controls, and WRAP that includes 64 patients and 200 controls; as for the speech length, Orozco-Arroyave Database comprises over 150 hours of speech divided in three sections: repetition of Spanish vowels, repetition of sentences and spontaneous speech; as for the media type, 9 of 26 databases and subdatabases have video records, and one was supported by brain scans.

The number of citations illustrates the influence of those databases, which is most critical for corpora development. As shown in table 2, AphasiaBank is most frequently cited by researchers, followed by WRAP and DAIC-WOZ. AphasiaBank is one of the pioneering database projects in language disabilities of patients with mental disorder, in which the Cantonese AphasiaBank is extending rapidly, representing a convincing successful model for Chinese database developers.

The three databases (the Orozco-Arroyave Database, the Cambridge Cookie-Theft Corpus and the DementiaBank) apply multiple methods in data recordings, including biological indexes and brain scanning, or they may cover a wide range of mental diseases such as anxiety, dementia, depression and PTSD. One common feature of them is the large scale of participants, which is an indispensable necessity for a reliable and scientific research study.

Among the overall nine independent databases, AphasiaBank and DementiaBank serve as pioneering resources in neurocognitive areas that may well provide references for future studies on mental disorder, especially the Cantonese AphasiaBank Database, which is the only one recorded in Chinese. Therefore, there is a lack of existing corpora in Asia, particularly in the LMICs, while Asian scholars in relevant fields are in great demand to apply corpora or databases containing transcriptions of patients with mental disorder in their own languages.

Limitations

The construction of speech corpora for mental disorders is an interdisciplinary task, which involves many fields such as neurocognitive science, linguistics and speech technology. Therefore, the existing corpora are yet to be further completed with advanced technological help, while the insufficiency in language diversity and disease comprehensiveness requests the effort of more contributors, especially those from Asian countries.

In general, the participants included in those corpora are restricted to a limited number under 40 in the majority of the studies, while those above 50 are rarely available. This requires the enlargement of samples. In addition, the corpora discussed in this article mostly focus on patients with neurocognitive illness, such as aphasia, stutter, and brain injuries, which reveal explicit signs of language disabilities in learning and delivering verbal or non-verbal information. There is an inadequate number of existing corpora for patients with mental disorders such as dementia, schizophrenia, depression, autism and other mental diseases that may also reveal abnormalities in emotion, language usage, or behavior in communication, which should be taken into consideration. Patients with mental disorders, though they might not be suffering from articulatory disabilities, are still likely to deliver symptoms of abnormal communication behaviours, which can be employed for early detection and timely treatment.

Implications

As mentioned above, researchers should take into account more types of mental illness including dementia, schizophrenia, depression, and so forth, while the languages included can also be more diversified by supplementing Chinese and other languages from LMICs or even their local dialects for a comprehensive coverage of data collection.

As for methods, the process of data collection is also shifting from unimodal to multimodal, that is, the media types are gradually changing from audio to video, and even with the assistance of brainimaging and electrophysiological techniques. This trend is explicitly revealed in the application of clinical trials. In future studies, more technological supports, such as the eye-tracking and the functional MRI, are highly recommended in developing a more advanced multimodal speech database.

The development of multimodal corpora for mentally impaired patients can provide great insights for both academic research and clinical applications. Research work using comprehensive corpora should by all means be grounded on an interdisciplinary basis. With joint efforts from various study domains, we may enrich the research paradigms by comparing the behavioral and neurological processing features of mentally impaired patients and healthy controls, thus further shedding light on the mechanism of speech production and perception. On the other hand, the constructed databases serve as important resources for developing products of artificial

intelligence, which establish a solid foundation for clinical diagnosis, rehabilitation and treatment in today's era of big data.

Acknowledgements We thank all reviewers and editors for their insightful suggestions and valuable guidance in the improvement of this review.

Contributors CBL has suggested and proposed the research topic and potential databases for the study. HWD has directed the research background, significance and discussion based on the result. YLL has done the research procedure and result analysis of the study. YL has helped in the literature search and screening.

Funding This study was supported by grants from the major project of National Social Science Foundation of China (18ZDA293) and the interdisciplinary programme of Shanghai Jiao Tong University (14JCZ03).

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No additional data are available.

REFERENCES

1 who.int [internet]. Geneva: World Health Organization; c2018 [cited 2018 Sept 26] *What is Mental Disorder?* in *World Health Organization*. Available: http://www.who.int/mental_health/management/en/
2 Herrman H, Saxena S, Moodie R. *Promoting mental health: concepts, emerging evidence, practice [Internet*. Geneva, Switzerland: World Health Organization, 2005.
3 Merriam-Webster [Internet]. *Corpus* in *Merriam-Webster Online*, 2018. Available: https://www.merriam-webster.com/dictionary/corpus [Accessed cited 2018 May 30].
4 Baxter AJ, Patton G, Scott KM, *et al*. Global epidemiology of mental disorders: what are we missing? *PLoS One* 2013;8:e65514–9.
5 MacWhinney B, Fromm D, Forbes M, *et al*. AphasiaBank: methods for studying discourse. *Aphasiology* 2011;25:1286–307.
6 ISRCTN registry [internet]. BMC: part of Springer nature; c2018. Available: http://www.isrctn.com/ [Accessed cited 2018 Sept 26].
7 LREC Conferences [internet]. The International Conference on language resources and evaluation; c2000, 2012. Available: http://www.lrec-conf.org/ [Accessed cited 2018 Sept 26].
8 WorldWideScience [internet]. WorldWideScience: the global science gateway. Available: https://worldwidescience.org/ [Accessed cited 2018 Sept 26].
9 Yu JM, Kim B, Lee K-M, *et al*. Symptomatic conceptualization of disorganized speech in patients with schizophrenia. *Korean Journal of Schizophrenia Research* 2015;18:51–8.
10 Becker JT, Boller F, Lopez OL, *et al*. The natural history of Alzheimer's disease. Description of Study cohort and accuracy of diagnosis. *Arch Neurol* 1994;51:585–94.
11 Kraljević JK, Hržica G, Lice K. CroDA: a Croatian discourse corpus of speakers with aphasia. *Croatian Review of Rehabilitation Research [Internet]* 2017;53:61–7.
12 Colin C, Le Meur C. *Adaptation Du projet AphasiaBank La langue Française: contribution pour une évaluation informatisée Du discours oral de patients aphasiques*. Toulouse: Université Paul Sabatier Toulouse III, 2016. (cited 2018 Sept 27).
13 Gaillard WD, Berl MM, Moore EN, *et al*. Atypical language in lesional and nonlesional complex partial epilepsy. *Neurology* 2007;69:1761–71.
14 Yairi E, Ambrose N. *Early childhood stuttering*. Austin: Pro Ed, 2005.
15 Westerhout E, Monachesi P. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, 2006, A pilot study for a Corpus of Dutch Aphasic Speech (CoDAS). Available: https://pdfs.semanticscholar.org/2e4f/6053452687acba279c9d23e2cefb5ee7011c.pdf
16 Williams C, Thwaites A, Buttery P, *et al*. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, 2010, The Cambridge Cookie-Theft Corpus: a corpus of directed and spontaneous speech of brain-damaged patients and healthy individuals. Available: https://csl.psychol.cam.ac.uk/publications/pdf/10_Williams_LREC.pdf
17 Varlokosta S, Stamouli S, Karasimos A. A Greek Corpus of Aphasic Discourse: collection, transcription, and annotation specifications. LREC Workshop: RaPID-2016 - 23rd of May 2016 - Portorož Slovenia, 2016. Available: http://www.ep.liu.se/ecp/128/003/ecp16128003.pdf
18 Orozco-Arroyave JR, Arias-Londoño JD, Vargas-Bonilla JF, *et al*. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, 2014, New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. Available:https://pdfs.semanticscholar.org/6a62/bc47de1bd1dea5113d7918f3f9bef521058b.pdf
19 Gratch J, Artstein R, Lucas G. The distress analysis interview corpus of human and computer interviews. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, 2014:3123–8.
20 Johnson SC. The Wisconsin Registry for Alzheimer's Prevention: a review of findings and current directions. *Alzheimers Dement* 2017;10:130–42.
21 Conroy P. Difficulty retrieving words (anomia) in people with relapsing-remitting multiple sclerosis (RR-MS). *ISRCTN registry*, 2018. Available: http://www.isrctn.com/ISRCTN54123111?q=Difficulty%20retrieving%20words%20(anomia)%20in%20people%20with%20relapsing-remitting%20multiple%20sclerosis%20&filters=&sort=&offset=1&totalResults=1&page=1&pageSize=10&searchType=basic-search
22 Chakrabortyy D, Yang Z, Tahir Y, *et al*. Prediction of Negative Symptoms of Schizophrenia from Emotion Related Low-Level Speech Signals. *IEEE SigPort*, 2018. Available: http://sigport.org/2428
23 Andreasen NC, Nancy CA. Negative symptoms in schizophrenia. Definition and reliability. *Arch Gen Psychiatry* 1982;39:784–8.
24 Patel D, Yaminiy BK, Meera SS, *et al*. Comparison of speech tasks for automatic classification of patients with amyotrophic lateral sclerosis and healthy subjects. *IEEE SigPort*, 2018. Available: http://sigport.org/3158
25 Volkmer A. Better conversations with primary progressive aphasia (BCPPA). *ISRCTN registry*, 2018. Available: http://www.isrctn.com/ISRCTN10148247?q=mental%20disorder%20language&filters=&sort=&offset=34&totalResults=462&page=4&pageSize=10&searchType=basic-search
26 Beitchman JH, Brownlie EB, Bao L. Age 31 mental health outcomes of childhood language and speech disorders. *Journal of the American Academy of Child & Adolescent Psychiatry* 2014;53:1102–10.
27 Dudy S, Asgari M, Kain A. Pronunciation analysis for children with speech sound disorders. *Conf Proc IEEE Eng Med Biol Soc* 2015:5573–6.
28 Goodglass H, Kaplan E, Barresi B. *The Boston Diagnostic Aphasia Examination*. 3rd. Philadelphia, PA: Lippincott, 2001.
29 Boschi V, Catricalà E, Consonni M, *et al*. Connected speech in neurodegenerative language disorders: a review. *Front Psychol* 2017;8.
30 Davy W, Travis JA, Laura W, *et al*. Automatic prediction of linguistic decline in writings of subjects with degenerative dementia. 2016 conference of the North American chapter of the Association for computational linguistics: human language technologies, NAACL Hlt. *Association for Computational Linguistics* 2016;2016:1198–207.
31 Kertesz A, Kintsch W. *The Western aphasia battery*. New York, NY: Grune and Stratton, 1982.
32 Machado T, Brandão L, Alice M, *et al*. Alzheimer's disease: cognition and picture-based narrative discourse. *CEFAC* 2014;16:1168–76.
33 Silveira G, Mansur LL. Analysis of prototypical narratives produced by aphasic individuals and cognitively healthy subjects. *Dement neuropsychol* 2015;9:279–84.
34 Fraser KC, Meltzer JA, Rudzicz F. Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *JAD* 2016;49:407–22.
35 Orimaye SO, Wong JS-M, Golden KJ, *et al*. Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. *BMC Bioinformatics* 2017;18.
36 Masrani V, Murray G, Field TS, *et al*. Domain Adaptation for Detecting Mild Cognitive Impairment. In: Mouhoub M, Langlais P, eds. *Advances in Artificial Intelligence. AI . Lecture Notes in Computer Science*. 10233. Cham: Springer, 2017.
37 Boyle M. Stability of Word-Retrieval errors with the AphasiaBank stimuli. *Am J Speech Lang Pathol* 2015;24:S953–S960.
38 Ellis C, Peach RK. Racial-Ethnic differences in word fluency and auditory comprehension among persons with poststroke aphasia. *Archives of Physical Medicine and Rehabilitation* 2017;98:681–6.

39 Mueller K, Koscik R, Hermann B, *et al*. Declines in connected language are associated with very early mild cognitive impairment: results from the Wisconsin Registry for Alzheimer's prevention. *Front Aging Neurosci* 2017;9:1–14.
40 Dham S, Sharma A, Abhinav D. Depression Scale Recognition from Audio, Visual and Text Analysis. *CoRR*, 2017, *abs*/1709.05865. Available: https://arxiv.org/abs/1709.05865v1
41 Cummins N, Vlasenko B, Sagha H, *et al*. Enhancing Speech-Based Depression Detection Through Gender Dependent Vowel-Level Formant Features. In: Teije A, Popow C, Holmes J, *et al*, eds. *Artificial intelligence in medicine. AIME 2017. Lecture Notes in computer science, vol 10259*. Cham: Springer, 2017.
42 Kong APH, Law SP, ASY L. The construction of a corpus of Cantonese-aphasic discourse: A preliminary report. In: *Poster presented at the American speech Language-Hearing Association (ASHA) convention*. New Orleans, LA, 2009.
43 Rahman A, Divan G, Hamdani SU, *et al*. Effectiveness of the parent-mediated intervention for children with autism spectrum disorder in South Asia in India and Pakistan (PASS): a randomised controlled trial. *Lancet Psychiatry* 2016;3:128–36.
44 Bersudsky Y, Fine J, Gorjaltsan I, *et al*. Schizophrenia and second language acquisition. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 2005;29:535–42.
45 Yu R, Wu R, Tai C, *et al*. Neuropsychological profile in patients with early stage of Parkinson's disease in Taiwan. *Elsevier: Parkinsonism Relat Disord* 2012;18:1067–72.
46 Sah W, Torng P. Production of mental state terms in narratives of Mandarin-speaking children with autism spectrum disorder. *Clin Linguist Phon* 2016;8:1–19.
47 Wu G, Xu X, Gu Y, *et al*. An overview on clinical speech disorders for dementia. *Contemp Ling* 2014;16:452–65.
48 Boersma P, Weenink D. PRAAT [Computer software]. Available: http://www.praat.org
49 Hellwig B. EUDICO linguistic Annotator (ELAN) version 5.3.0 manual, 2018. Available: https://www.mpi.nl/corpus/html/elan/ [Accessed cited 2018 Sept 15].
50 Fung HK-H, Ho GP-C, Kong APH, *et al*. Applying main concept analysis (MCA) to analyze spoken discourse by Cantonese speakers with aphasia and unimpaired individuals. *Front Hum Neurosci* 2017;11. Conference Abstract: *Academy of Aphasia 55th Annual Meeting*.
51 Nicholas LE, Brookshire RH. Presence, completeness, and accuracy of main concepts in the connected speech of non-brain-damaged adults and adults with aphasia. *J Speech Lang Hear Res* 1995;38:145–56.
52 Lee T, Kong APH. Automatic speech assessment for aphasic patients based on syllable-level embedding and supra-segmental duration features. *IEEE SigPort*, 2018. Available: http://sigport.org/257