INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Towards the Speech Features of Early-stage Dementia: Design and Application of the Mandarin Elderly Cognitive Speech Database

Tianqi Wang[1], *Quanlei Yan*[1], *Jingshen Pan*[1], *Feiqi Zhu*[2], *Rongfeng Su*[1], *Yi Guo*[3], *Lan Wang*[1], *Nan Yan*[1]

[1]CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[2]Department of Neurology, Shenzhen Luohu People's Hospital, Shenzhen, China
[3]Department of Neurology, Shenzhen People's Hospital, Shenzhen, China
zfqzsu2004@aliyun.com, nan.yan@siat.ac.cn, lan.wang@siat.ac.cn

Abstract

Speech and language features have been proven to be useful for the detection of neurodegenerative diseases, such as Alzheimer's disease (AD), and its prodromal stage, mild cognitive impairment (MCI). Unfortunately, high-quality speech database remains scarce, which limit its application in automatic screening and assessment of early dementia in clinical practice. To bridge this gap, the present study aimed to design a speech database of Chinese elderly with intact cognition and MCI, named "Mandarin Elderly Cognitive Speech Database" (MECSD). The database consists of 110 hours of speech recordings from 85 native speakers of Mandarin Chinese (age range = 55 – 85 years), including 20 participants with MCI and 65 healthy controls. Manually transcribed materials with temporal information were also included in this database. Nine tasks, involving conventional test batteries and connected speech productions, were used to obtain speech samples, producing a total of 8563 sentences and 49841 words. Details concerning the design of the database, together with our preliminary findings applying automatic speech recognition (ASR), were reported in this study. The MECSD will provide researchers with access to a large shared database that can facilitate hypothesis testing in the study of early-stage dementia.

Index Terms: early-stage dementia, mild cognitive impairment, speech database, Mandarin corpus

1. Introduction

Alzheimer's disease (AD) is one of the most common type of neurodegenerative diseases, affecting over 7% of the population aged 60 years or more [1]. It is estimated that the total number of individuals with dementia will reach 82 million worldwide in 2030, accounting for 60 - 70% of AD cases [2]. Of particular interest is the fact that a significant proportion of the elderly suffered mild cognitive impairment (MCI) [3], a prodromal stage of AD, and approximately 50% of individuals in this spectrum developed dementia over a 5-year period [4]. Therefore, early identification of cognitive decline associated with MCI is of vital importance, as this preclinical period could provide a valuable time window for drug development, risk assessment, and prevention [5-7].

In recent years, increasing interest is directed towards the exploration of non-invasive biomarkers for the identification of early-stage dementia. Complex cognitive functions, such as language, are natural candidates for this. In particular, a well-documented literature has suggested that alterations in speech and language might be one of the earliest signs of cognitive decline. Therefore, spoken language has become a research tool, as well as a diagnostic resource, towards the cognitive status of individuals with psychometric evidence of MCI.

Realizing automatic or semiautomatic analysis of speech samples is another shared goal among researchers, as it will help create low-cost tools for early detection of cognitive decline among large sections of the population [8]. With the identification of acoustic, lexical, semantic, syntactic, and pragmatic features associated with MCI [9-11], a series of parameters could be potentially applied to recognize, classify, and describe early-stage dementia. In that sense, the collection of high-quality speech samples is in desperate need.

To date, very limited open databases were designed for the study of communication in dementia. One is the DementiaBank corpus [12], which includes speech samples from patients with "possible" or "probable" AD. The data were collected between 1983 and 1988 as part of the Alzheimer Research Program at the University of Pittsburgh, and the current database comprises speech samples in English, German, Spanish, and Mainland and Taiwanese Mandarin. One limitation of this data set is the unmatched demographic features in patient and control groups, where subjects in the AD group tend be older ($p < 0.01$) and less educated ($p < 0.01$) [13]. The other limitation derives from the limited variety of tasks used to elicit speech production. In the DementiaBank corpus, both situational picture description from the Boston Diagnostic Aphasia Examination [14] (i.e., *the Cookie Theft*) and the speech fluency tests from conventional neuropsychological batteries were applied to obtain connected and isolated speech samples. Interestingly, a recent systematic review on the connected speech of neurodegenerative disorders suggested that semi- or unstructured speech productions were superior approaches to obtain speech samples in more natural settings. Further advantage of this task manifest itself in the discourse and pragmatic levels of processing [15]. Last but certainly not least, speech samples in this widely used corpus were obtained from participants with the moderate to severe AD, which may not be effective in detecting very early stage of cognitive decline.

In the absence of such a speech database which captures subtle language deficits associated with MCI in both clinical and everyday communication context, the "Mandarin Elderly Cognitive Speech Database" (MECSD) was designed to bridge this gap. The central goal of the MECSD project is to provide researchers, especially those who were interested in Mandarin

http://dx.doi.org/10.21437/Interspeech.2019-2453

Chinese, with access to a large, shared database that can facilitate hypothesis testing in the study of early-stage dementia. The collection of MECSD materials began in 2018, and by now the database had grown to include 20 participants with MCI (hereafter as MCIs) and 65 cognitively intact controls (hereafter as controls).

In the present study, we introduced how MECSD was constructed, and presented sample analyses using large vocabulary continuous speech recognition (LVCSR) technique. It is worth mentioning that MECSD represents the first speech database of Mandarin Chinese which comprises speech samples produced in a variety of task settings. Applying the data set for speech and language analysis will carry important implications to both research and clinical practice towards the understanding of specific features associated with early-stage dementia.

2. Data Collection

2.1. Subject selection

Speech samples produced by a total of 85 individuals including 20 MCIs, and 65 controls, were collected for MECSD. The MCI participants recruited from Shenzhen Luohu People's Hospital received comprehensive neuropsychological evaluations and were assigned to the MCI group based on a history of cognitive decline and the results of the mental status examination (i.e., Montreal Cognitive Assessment, MoCA; and Mini-Mental State Examination, MMSE; Beijing Version). Participants having no history of psychiatric issues or neurological disorders were also recruited as controls. The inclusion and exclusion criteria are listed in Table 1.

Table 1: *Inclusion and exclusion criteria for participant enrollment in MCI and control groups*

MCI	Control
Inclusion criteria:	
- MoCA ≥ 18	- MoCA ≥ 26
- MMSE ≥ 18	- MMSE ≥ 24
- Age between 55 and 85	
- At least primary school education level. No maximum education level limit.	
- Sufficient Chinese language skills to complete all testing.	
- Sufficient vision and hearing to engage in all testing.	
Exclusion criteria [16]:	
- Diseases associated with dementia, such as AD, Parkinson's disease, ischemic vascular dementia.	
- Current or history of major psychiatric diseases, such as depression and schizophrenia.	
- Significant diseases of central nervous system such as brain tumor, seizure disorder, subdural hematoma, cranial arteritis.	
- Current alcohol or substance abuse.	
- Current or history of taking neuroleptic medication.	

Table 2 reports the demographic information of the current sample set, as well as the row scores of MoCA and MMSE that are corrected for age and education.

Table 2: *Demographic information (mean, SD, and range) of subjects in the current MECSD data set*

	MCI **N = 20**	**Normal** **N = 65**
Gender (F/M)	12/8	35/30
Age (years)		
- **Mean**	65.85	67.71
- **SD**	5.53	5.77
- **Range**	(60, 82)	(56, 82)
MoCA (30)		
- **Mean**	23.7	28.0
- **SD**	2.2	1.3
- **Range**	(18, 27)	(25, 30)
MMSE (30)		
- **Mean**	25.4	29.2
- **SD**	2.7	1.0
- **Range**	(20, 30)	(26, 30)

2.2. Speech and language tasks

A major contribution of MECSD is that we designed up to nine tasks to gauge speech and language performance of the two groups of participants, which include conventional test batteries highlighting functions in verbal fluency and naming (e.g., Boston Naming Test), as well as the structured and unstructured connected speech production tasks (e.g., situational picture description and spontaneous self-introduction), which elicit speech productions in more natural settings. In addition, we also designed two tasks to examine phonological working memory and semantic memory, allowing for post-analysis of the audio recordings to investigate the effect of neuropsychiatric symptoms on the speech and language parameters.

Respective descriptions of the nine tasks are reported in Table 3.

In all connected speech production tasks (i.e., SI and PD), the examiner was allowed to encourage the participant to keep going if they produced a very limited speech output. All the instructions and interventions were included in recording session.

2.3. Recording protocol

The vocal tasks were recorded in a noise-attenuated room using an external sound mixer (E-MU 0404 USB Audio/MIDI Interface, Creative) connected to a professional microphone (MK4, Sennheiser) which was placed 10 cm from the lips. The samples were digitized with 44100 Hz sampling rate and 16-bit/sample resolution.

2.4. Data segmentation

Audio files were segmented after collection. Excessive silence exceeding task duration was removed. Utterances were further segmented into sentences with about 0.2 seconds at the start and end of each sentence. It should be noted that MECSD provides access to a parallel set of audio files, both at sentence and task level. The LVCSR was performed at sentence level, while multidimensional linguistic features (e.g. lexical, semantic, syntactic, and acoustic parameters) could be analyzed at task level.

Table 3: *Descriptions of the tasks used to obtain speech and language samples for MECSD*

Task 1: Self-introduction (SI)
The purpose of SI is to elicit spontaneous (unstructured) speech production by asking predefined open-ended questions, such as "Tell me about your career/family/hobbies". As a warming-up section of the test, a brief self-introduction could also get participants ready for the rest of the test. The duration of SI is one minute.

Task 2: Picture description (PD)
Three pictures, including the most widely used "Cookie Theft" picture description task from the Boston Diagnostic Aphasia Examination (BDAE) [14], are used to elicit structured speech production. Two other picture stimuli are chosen from the Western Aphasia Battery (WAB) [17] During the stimuli presentation, participants are asked to describe everything they see in the picture. To allow comparison between structured and unstructured connected speech production, each picture stimulus is presented for one minute.

Task 3: Speech Fluency (SF)
SF task is designed to examine verbal functioning [18], in which participants are given one min to produce as many unique words as possible within a semantic category (fruits, animals, and place names in our task).

Task 4: Picture Naming (PN)
Thirty-six line-drawing stimuli are chosen from the Snodgrass and Vanderwart picture set [19]. A norming study is conducted to ensure name agreement, image agreement, and familiarity among another group of Chinese elderly. In this task, participants are asked to name each picture as quickly as accurately as possible.

Task 5: Sentence Repetition (SR)
Fourteen sentences with increasing number of words and syntactic complexity are orally presented to the participant one after another. Participants are required to repeat each sentence immediately after its presentation.

Task 6: Poem Reading (PR)
Six ancient Chinese poems are presented to the participant one after another. Participants are asked to read each poem aloud.

Task 7: Articulation (AR)
Participants are asked to articulate syllables "pa-ta-ka" in a row for three times.

Task 8: Span Task (ST)
The ST task derives from the examination of phonological short-term memory in [20], in which participants are asked to recall visually presented sequences of an increasing number (3 - 5) of monosyllabic Chinese morphemes having either the same tone or different tones. The number of correctly-recalled items is obtained as a measure of their short-term memory span.

Task 9: Semantic Matching (SM)
The SM task derives from the Pyramid & Palm Tree Test (PPTT), in which participants are required to perform Picture-to-Picture Matching based on the judgement of whether the two items presented are used together (e.g., hammer and nail), or share the same function (e.g., radio and CD player).

2.5. Transcription protocol

Speech samples were manually transcribed at the word level by three researchers (TW, JP, and QY) using TextGrid in *Praat*. In order to ensure consistency among transcribers, the TalkBank Codes for Human Analysis of Transcripts (CHAT) [21] protocol was imposed. Unrelated utterances such as questions about the task or conversations with the examiner were ignored. Paralinguistic phenomena such as filled pauses (e.g., "uh", "mm", "er", "ah", "zhege", "nage" ["this" and "that" in Chinese]), disfluencies (e.g., false start, hesitation, stuttering), and non-verbal phenomena (e.g., coughing, throat clearing, laughs, inspirations) were also annotated, together with temporal information.

3. Descriptive Statistics of MECSD

MECSD includes 15.4 hours (924.15 minutes) of isolated and connected speech production, resulting in a total of 8563 sentences and 49841 words. Table 4 and 5 reports descriptive statistics of speech samples in each single task.

Table 4: *The duration of speech samples in each task (Unit: minute)*

Task	MCI	Normal
SI	22.77	83.71
PD	74.75	256.17
SF	60.89	198.77
SR	22.92	68.19
PR	33.77	102.21
Total	215.10	709.05

Table 5: *Number of sentences, words, and unique words produced by MCIs and controls*

	MCI	Normal
Sentences	2050	6513
Words	11722	38119
Unique words	1195	3496

4. Pilot ASR Experiment

MECSD can be applied to speech signal analysis, speaker recognition and speech recognition. The Automatic Speech Recognition (ASR) system has been proposed to identify dementia in early stages [22]. To reveal differences of speech production between MCIs and healthy controls in MECSD, a pilot ASR experiment was conducted. The results could serve as a reference for future studies.

4.1. Experimental setup

Since it is difficult to use the limited speech data collected from Chinese elderly persons to construct ASR systems for LVCSR purpose, the DARPA GALE Mandarin broadcast news data (totally 562 hours) were used to obtain the ASR systems. Among these, 90% of the broadcast data were used as the training set, while the rest were used as the development set. The MECSD set described in Section 3 was used as the test set. The whole test set can be divided into two subsets: MCI and Normal. In each subset, only the speech data from the PD and SI modules described in Section 2.2 were used for testing, since they were continuous speech data. An interpolated tri-gram language model (LM) was used for decoding. It was obtained from four sub-LMs trained on the GIGAWORD Mandarin

Chinese transcription, TDT4 Mandarin Chinese transcription, HGSD prompts, and online corpus respectively.

4.2. ASR systems

Four ASR systems were used to evaluate the dataset proposed in this study. They were described as follows:

- **GMM-HMM system:** The GMM-HMM baseline was maximum likelihood (ML) trained using the HTK toolkit. The acoustic inputs were 48-dimensional HLDA projected PLP+Kaldi pitch features. They were obtained as followed: 52-dimensional PLP features (static, first, second and third order differentials) were first augmented with 12-dimnesional Kladi pitch features (static, first, second and third order differentials) to form a 64-dimensional features; a HLDA transform (from 64 dimensions to 48 dimensions) was then applied on such features. A total of 6384 tied tri-phone states with 24 Gaussian components per state were used.
- **GMM-HMM-MLLR system:** In order to improve the performance of the GMM-HMM baseline system, the unsupervised, speaker-level MLLR based adaptation technique was used in testing.
- **Hybrid DNN system:** The structure of the hybrid DNN system contained 6 hidden layers with sigmoid activation. Each hidden layer contained 2048 nodes. A context window splicing by 11 successive frames of 48-dimensional HLDA projected PLP+Kaldi pitch features was used as the acoustic inputs. The state-level alignment information for the DNN training was produced using the baseline GMM-HMM system.
- **Hybrid DNN-SAT system:** The constrained maximum likelihood linear regression (CMLLR) based speaker adaptive training (SAT) was performed on the hybrid DNN system described above. We first estimated a CMLLR transform for each speaker using the baseline GMM-HMM system, and then applied the estimated speaker-level CMLLR transform on the 48-dimensional HLDA projected PLP+Kaldi pitch features. 11 consecutive frames of the resulting features were used as the neural network inputs. The hybrid DNN-SAT system use the same architecture and training criterion in the hybrid DNN system. The CMLLR transformation of the test sets were estimated on the output results of the Hybrid DNN system without SAT.

4.3. Experiment results

Table 6 shows the system performance evaluated by the character error rate (CER) in the PD and SI module. Using the speaker adaptation technique, CER reductions of 8%-16% and 12%-23% were obtained over the GMM-HMM and Hybrid DNN systems respectively. This indicates that the system performance can be significantly improved by removing the variability introduced by speakers. The results also revealed that the ASR system performed worse on the MCI subset, suggesting that patients, even in the prodromal stage of AD, are associated with impairment in speech production.

5. Conclusion

MECSD, a high-quality collection of speech data from 85 elderly Mandarin Chinese speakers with normal cognition and mild cognitive impairment, provides researchers with access to a large shared database that can facilitate hypothesis testing in the study of speech features associated with cognitive decline. The present study specified the database design, with a pilot ASR experiment aiming at revealing differences in speech production between MCIs and healthy controls.

Table 6: *CER performance of GMM-HMM and Hybrid DNN systems*

Normal Subset					
ASR	**PD 01**	**PD 02**	**PD 03**	**SI**	**Overall**
GMM-HMM	37.79	42.22	43.43	36.86	39.50
GMM-HMM-MLLR	31.92	35.55	36.69	30.81	33.28
Hybrid DNN	24.74	27.44	28.96	24.68	26.03
Hybrid DNN-SAT	21.27	23.04	24.56	20.02	21.95
MCI Subset					
ASR	**PD 01**	**PD 02**	**PD 03**	**SI**	**Overall**
GMM-HMM	52.21	59.52	53.67	56.58	55.38
GMM-HMM-MLLR	47.28	53.90	49.49	51.40	50.37
Hybrid DNN	42.86	49.69	45.76	50.68	47.03
Hybrid DNN-SAT	34.01	40.39	38.90	38.65	37.70

Note: PD 01-03 represents the three picture description tasks.

6. Acknowledgement

This study was jointly supported by a grant from National Natural Science Foundation of China (61771461 and U1736202) and Shenzhen Fundamental Research Program (JCYJ20170413161611534 and JCYJ20150330102401089).

7. References

[1] B. Duthey, *Background Paper 6.11: Alzheimer Disease and other Dementias*. WHO, 2013.

[2] (2017, Dec 12). *Dementia*. Available: https://www.who.int/en/news-room/fact-sheets/detail/dementia

[3] R. C. Petersen, G. E. Smith, S. C. Waring, R. J. Ivnik, E. G. Tangalos, and E. Kokmen, "Mild cognitive impairment: Clinical characterization and outcome," *Archives of Neurology,* vol. 56, no. 3, pp. 303-308, Mar 1999.

[4] S. Gauthier *et al.*, "Mild cognitive impairment," *Lancet,* vol. 367, no. 9518, pp. 1262-1270, Apr 15 2006.

[5] L. Calzà *et al.*, "Should we screen for cognitive decline and dementia?," *Maturitas,* vol. 82, no. 1, pp. 28-35, 2015.

[6] S. Epelbaum *et al.*, "Preclinical Alzheimer's disease: A systematic review of the cohorts underlying the concept," *Alzheimers & Dementia,* vol. 13, no. 4, pp. 454-467, Apr 2017.

[7] K. Ritchie *et al.*, "Recommended cognitive outcomes in preclinical Alzheimer's disease: Consensus statement from the European Prevention of Alzheimer's Dementia project," *Alzheimers & Dementia,* vol. 13, no. 2, pp. 186-195, Feb 2017.

[8] D. Beltrami, G. Gagliardi, R. R. Favretti, E. Ghidoni, F. Tamburini, and L. Calza, "Speech analysis by natural language processing techniques: A possible tool for very early detection of cognitive decline?," *Frontiers in Aging Neuroscience,* vol. 10, Nov 13 2018.

[9] B. Roark, M. Mitchell, J. P. Hosom, K. Hollingshead, and J. Kaye, "Spoken language derived measures for detecting mild cognitive impairment," *IEEE Transactions on Audio Speech and Language Processing,* vol. 19, no. 7, pp. 2081-2090, Sep 2011.

[10] A. Satt *et al.*, "Evaluation of speech-based protocol for detection of early-stage dementia," *14th Annual Conference of the International Speech Communication Association (Interspeech 2013), Vols 1-5,* pp. 1691-1695, 2013.

[11] S. Ahmed, A. M. F. Haigh, C. A. de Jager, and P. Garrard, "Connected speech as a marker of disease progression in autopsy-

proven Alzheimer's disease," *Brain,* vol. 136, pp. 3727-3737, Dec 2013.

[12] B. MacWhinney, D. Fromm, M. Forbes, and A. Holland, "AphasiaBank: Methods for studying discourse," *Aphasiology,* vol. 25, no. 11, pp. 1286-1307, 2011.

[13] K. C. Fraser and G. Hirst, "Detecting semantic changes in Alzheimer's disease with vector space models," presented at the Language Resources and Evaluation Conference, Portoroz, Slovenia, 2016.

[14] H. Goodglass and E. Kaplan, *Boston diagnostic aphasia examination booklet*. Philadelphia, PA: Lea & Febiger, 1983.

[15] V. Boschi, E. Catricala, M. Consonni, C. Chesi, A. Moro, and S. F. Cappa, "Connected speech in neurodegenerative language disorders: A review," *Frontiers in Psychology,* vol. 8, pp. 1-21, Mar 6 2017.

[16] M. Asgari, J. Kaye, and H. Dodge, "Predicting mild cognitive impairment from spontaneous spoken utterances " *Alzheimers & Dementia* vol. 3, no. 2, pp. 219-228, 2017.

[17] A. Kertesz, *Western Aphasia Battery-Revised*. New York: Grune & Stratton, 2006.

[18] M. Lezak, D. Howieson, E. Bigler, and D. Tranel, *Neuropsychological Assessment*. New York, NY: Oxford University Press, 2012.

[19] J. G. Snodgrass and M. Vanderwart, "A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity," *Journal of Experimental Psychology: Human Learning and Memory,* vol. 6, no. 2, pp. 174-215, 1980.

[20] Y. Xu, "Depth of phonological recording in short-term memory," *Memory & Cognition,* vol. 19, no. 3, pp. 263-273, 1991.

[21] B. MacWhinney, *The CHILDES project: Tools for analyzing talk: Volume I: Transcription format and programs*. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

[22] G. Gosztolya, V. Vincze, L. Toth, M. Pakaski, J. Kalman, and I. Hoffmann, "Identifying Mild Cognitive Impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features," *Computer Speech and Language,* vol. 53, pp. 181-197, Jan 2019.