INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Integrated and Enhanced Pipeline System to Support Spoken Language Analytics for Screening Neurocognitive Disorders

Helen Meng[1,2,3], *Brian Mak*[4], *Man-Wai Mak*[5], *Helene Fung*[6], *Xianmin Gong*[2,6], *Timothy Kwok*[7,8,9], *Xunying Liu*[1], *Vincent Mok*[10,11,12,13], *Patrick Wong*[14,15], *Jean Woo*[7,9], *Xixin Wu*[2], *Ka Ho Wong*[1], *Sean Shensheng Xu*[5,16], *Naijun Zheng*[1], *Ranzo Huang*[4], *Jiawen Kang*[1], *Xiaoquan Ke*[5], *Junan Li*[3], *Jinchao Li*[1], *Yi Wang*[1]

[1–3,6–15]The Chinese University of Hong Kong, [4]The Hong Kong University of Science and Technology, [5]The Hong Kong Polytechnic University, [16]Shenzhen University, China

Abstract

This paper presents an enhanced pipeline system for automated screening of neurocognitive disorders, e.g. Alzheimer's Disease (AD), using spoken language technologies. To ensure local relevance, the pipeline is applied to two-way interactions between clinical assessors and older adult participants in spoken Cantonese, the predominant language used in Hong Kong. The pipeline includes: (i) Speaker diarization using speaker-turn-aware scoring to capture the temporal structure of conversations. (ii) ASR using XLS-R wav2vec 2.0 models further pre-trained on Cantonese speech data and fine-tuned. (iii) Language modelling using RoBERTa with further fine-tuning. (iv) AD screening with neural network classification. A reference benchmark is obtained using the ADReSS corpus where no diarization is needed, and the partial pipeline attained a competitive detection accuracy of 87.5%.

Index Terms: diarization, speech recognition, NCD detection, neurocognitive disorder, dementia

1. Introduction

In recent years, we have seen a proliferation of research interests in using spoken language technologies to screen Neurocognitive Disorders (NCDs), e.g., Alzheimer's Disease. This paper presents an overview of our integrated and enhanced pipeline system for the automated NCD screening. The pipeline includes speaker diarization, automatic speech recognition (ASR), language modelling, and finally a neural network classifier. To ensure local relevance, our pipeline is applied to two-way spoken interactions in Cantonese Chinese between an assessor and an older adult participant. The spoken interactions cover standardized neurocognitive tests and a specially designed dialog task that captures everyday cognitive competence. Speaker diarization of the two-way spoken interactions applies a speaker-turn-aware scoring approach to capture the temporal structure of conversations. ASR uses XLS-R wav2vec 2.0 models further pre-trained on Cantonese speech data and fine-tuned. We also apply the RoBERTa language model, further fine-tuned for the task, together with model combination for robustness. NCD screening is done by neural network classification. We obtained a reference benchmark using the publicly available ADReSS corpus for which no diarization is needed – the (partial) pipeline attains state-of-the-art AD detection accuracy of 87.5%. This paper will further report on the pipeline system's performance for the Cantonese speech corpus.

*[1]: Dept. of Systems Engineering & Engineering Management, [2]: Stanley Ho Big Data Decision Analytics Research Centre, [3]: Centre for Perceptual & Interactive Intelligence, [4]: Dept. of Computer Science & Engineering, [5]: Dept. of Electronic & Information Engineering, [6]: Dept. of Psychology, [7]: Dept. of Medicine & Therapeutics, [8]: Jockey Club Centre for Osteoporosis Care & Control, [9]: Jockey Club Institute of Aging, [10]: Division of Neurology, Dept. of Medicine & Therapeutics, [11]: Margaret K. L. Cheung Research Centre for Management of Parkinsonism, [12]: Li Ka Shing Institute of Health Sciences, [13]: Gerald Choa Neuroscience Institute, [14]: Dept. of Linguistics & Modern Languages, [15]: Brain & Mind Institute, [16]: School of Biomedical Engineering,

This project is partially supported by the HKSAR Research Grants Council (Project No. T45-407/19N).

2. Related Work

As a key domain of cognitive functioning, language ability could be impacted at an early stage of NCD, which causes symptoms such as temporal disfluency, reduced vocabulary coverage, and difficulties in word finding and retrieval [1, 2]. These findings provide a theoretical basis for utilizing spoken language as a means of detecting NCD, which could potentially offer a non-intrusive, objective, scalable and cost-effective solution for widespread NCD screening. A standard pipeline for conversation-based screening consists of a speaker diarization module for segmenting speech from older adult participants, an ASR module for generating transcripts, and classifiers based on features extracted from speech signal or recognized transcripts. The TDNNs/PLDA/AHC framework has been widely used in speaker diarization systems [3]. In this framework, time-delay neural networks (TDNNs) are utilized for extracting speaker embeddings, probabilistic linear discriminant analysis (PLDA) is applied to calculate similarity measures between all pairs of segments, and agglomerative hierarchical clustering (AHC) is employed for clustering and merging. Recent studies have focused on improving the diarization performance by enhancing the front-end representations [4] and back-end scoring [5, 6]. Previous research in feature engineering [7–11] have explored hand-crafted acoustic and linguistic features that are related to NCD, such as ComParE [12], eGeMAPS [13], part-of-speech (POS) [14] and syntactic complexity [15]. Their results show that most of the top-ranking features for NCD detection include dysfluency, verbal complexity and semantic richness. Recently, with advancements in self-supervised learning (SSL) approaches, the developments of NCD detection have been shifting from these low-level raw features to high-level pretrained embeddings, particularly in the domains of speech (e.g., VGGish [16], Wav2Vec 2.0 [17], OpenL3 [18], Whisper [19]) and text (e.g., BERT [11,20–22], ERNIE [21], Glove [23]). To effectively utilize these features, researchers proposed classifiers such as SVM and MLP [11], ensemble models such as voting [24], and long-form dependency cognitive modeling such as Attentive pooling [19].

Most previous work on NCD detection showed that the linguistic features obtained from transcriptions are generally more important than acoustic ones. To achieve automation, the devel-

10.21437/Interspeech.2023-2249

opment of automatic speech recognition (ASR) systems plays an essential role. Due to a large mismatch between young and elderly speech, as well as rich and low-resource languages, research efforts have been devoted to the adaptation of the ASR systems [25–27]. Interestingly, not only are the features from transcriptions, but the performance disparity in ASR between distinct diagnostic groups (i.e., healthy individuals versus those with NCD) could also help NCD detection [25].

3. Corpora

This work uses the publicly available English ADReSS dataset [9] for bench-marking the performance of the pipeline system. ADReSS is a subset of the Pitt Corpus in the DementiaBank dataset [28], and consists of 156 speech samples and associated transcripts from non-AD (35 male, 43 female) and AD (35 male, 43 female) English-speaking participants for the Cookie Theft picture description task, and is divided into standard training (108 participants, about 2 hours) and test (48 participants, about 1 hour) sets that with balanced distributions of age, gender and disease condition.

In addition, we also applied the pipeline to a Cantonese corpus that we have designed and collected, with the consideration that Cantonese Chinese is the major dialect used in Hong Kong, especially for our older adult population. The corpus is named MARVEL (Cognitive Assessment Using Machine LeArning empoweRed VoicE anaLysis). Participant inclusion criteria are: (i) aged 60 or above; (ii) sufficient Cantonese language skills (listening and speaking) and (iii) sufficient vision and hearing ability (where glasses and hearing aids are acceptable) to complete all cognitive tests.

MARVEL contains speech recordings of spoken interactions, where an assessor guides the participant through a comprehensive series of cognitive tests in a session. Each session requires an average of 1.5 hours, consisting of several cognitive tests, including *Alzheimer Disease Brief Information Interview* (AD8) [29,30], *Montreal Cognitive Assessment Hong Kong Version* (MoCA) [31] (a comprehensive NCD screening tool with the scores adjusted based on the age and education backgrounds of the participants), *Hong Kong List Learning Test* (HKLLT) [32] (a word list recall assessment), *Digit Span* (DS) [33] (a task that requires participants to repeat a series of numbers), *Modified Boston Naming Test* (mBNT) [34] (a picture naming task), *Logical Memory Story Telling* (LMST) [33], and *Rabbit Story* [35] (a narrative assessment), etc. We have also included a specially designed task, named the *Hong Kong Grocery Shopping Dialog Task* (HK-GSDT) [36], which combines spoken dialog with way-finding and memory recall of the shopping list.

Based on the speech recording of the cognitive tests, each participant will be classified by the clinicians into one of three groups: healthy controls, m(ild)-NCD and M(ajor)-NCD. Table 1 shows the number of participants in each group (totaling 585), while the project is ongoing and the numbers are expected to increase. The table also shows the number of participants whose speech data have been manually transcribed, which is a costly and laborious process, hence ASR techniques are leveraged to obtain automatic transcripts.

Orthographic transcription of the spontaneous Cantonese speech in MARVEL adopts the traditional Chinese character set (since this is the character set commonly used in Hong Kong). Colloquial words (e.g. "嘅"/ge3/ [possessive particle in colloquial Cantonese]) are also included. The average number of Chinese characters in the transcribed section of the participant's speech in a session is around 3600.

Table 1: *Number of participants classified into the three groups: healthy controls (HC), mild-NCD (m-NCD) and Major-NCD (M-NCD). "Trans." denotes manually transcribed speech data.*

	Train			Test		
	Trans.	Not Trans.	Total	Trans.	Not Trans.	Total
# of Participant	164	301	465	46	74	120
HC	76	222	298	22	54	76
m-NCD	69	71	140	19	15	34
M-NCD	19	8	27	5	5	10

The recordings from 585 participants have been validated. Division of data into train and test sets considers the balance across age (60-69, 70-75, and 76+), gender, health status, and available manual transcriptions. The duration of a recording ranges from 1.3 to 1.9 hours, with a mean of 1.5 hours. The summary is presented in Table 2.

Table 2: *The duration (hours) of train and test datasets, and transcribed (trans.) data. (*) Each session includes speech, silence and noisy parts.*

	Train			Test		
	Trans.	Not Trans.	Total	Trans.	Not Trans.	Total
Whole session*	126	582	707	35	152	187
Assessor speech	50	NIL	50	15	NIL	15
Participant speech	47	NIL	47	13	NIL	13

4. Speaker Diarization

Speaker diarization is performed by the first module in the pipeline. Since the raw speech recordings generally capture two-way dialogs between assessors and participants, the objective of speaker diarization is to enable extraction of participants' speech for further processing.

We adopted a clustering-based diarization approach, which mainly consists of an embedding extraction process, similarity measurement, clustering process and overlapped speech detection. Specifically, we applied a 40-dimensional Mel filter bank (Mel-FBank) with cepstral mean and variance normalization to obtain the fbank features. Next, a TDNN-LSTM voice activity detection (VAD) model trained on the MARVEL dataset was used to filter out non-speech segments, such as background noise. Based on the detected speech boundaries from the VAD model, the speech segments were sliced into sub-segments with a fixed-length window of 1.5 seconds and a sliding overlap of 1.25 seconds. Given the sub-segments, a 101-layer Res2Net model [37] pretrained on the VoxCeleb [38, 39] and CN-Celeb [40, 41] datasets was conducted to extract the speaker embedding. The speaker embedding was fed into a speaker-turn aware scoring model [6] to generate corresponding similarity scores, which were then organized into a similarity matrix. The similarity matrix contains both static and dynamic information across the conversation sub-segments by combining a pairwise similarity measure (e.g., Cosine) and LSTM-based similarities [5]. Initialized with the speaker-turn aware scoring matrix, the VBx algorithm [42] was applied to cluster all sub-segments within a conversation recording into different speaker groups.

Conventional diarization systems generally do not have special provision for overlapped speech that commonly occur in conversations. In this work, we applied a pyannote over-

lap detection model [43] trained on the Alimeeting [44] and AISHELL4 [45] dataset, it consists of SincNet convolution layers and LSTM layers that jointly train for VAD and OSD tasks. Overlapped speech is detected and since these segments contain the participant's speech, they are passed down the pipeline (albeit without speaker separation presently) for further processing.

To evaluate the speaker diarization module based on the MARVEL test set, we used the d-score toolkit,[1] and the results are shown in table 3. We observe that the overall diarization error rate (DER) is 5.22%. We further evaluate the miss rate and false alarm rate for assessors and participants separately. For example, to evaluate the performance on the participants, we replaced the diarization results for the assessors with ground-truth labels. Through comparison between the second and third rows in the table, we observed that the participant's speech is more challenging than the assessor's speech, as reflected especially by the miss rate. We attribute this to the higher variability in the participants' speech.

Table 3: *Diarization results (%) on the MARVEL test set. DER stands for diarization error rate.*

	Miss	False Alarm	Confusion	DER
Overall	2.47	1.70	1.05	5.22
- Assessor	1.25	1.30	-	-
- Participant	2.29	1.44	-	-

5. Automatic Speech Recognition

As manually transcribing the conversation speech recording is costly, we attempt to leverage the unlabeled speech data by pre-trained self-supervised speech representations, which has been shown to improve downstream ASR tasks in scenarios with limited labeled data [46]. We applied speaker diarization (discussed in the previous section) to the unlabeled parts of the training data from MARVEL, as well as another Cantonese elderly speech database named JCCOCC-MoCA [6]. Combining the manually labeled speech data with the unlabeled speech segments from both corpora yields totally 503 hours of pre-training data. This data is used to further pre-train and adapt the XLS-R model with 300M parameters [47], as a fast prototyping procedure. The adapted model is then used to create lexicon-based ASR systems, which are implemented in fairseq [48] with an integration of k2[2] for CTC training with alternative pronunciations and WFST-based decoding.

The decoding is conducted using an HLG graph based on a word 3-gram LM, and re-scored by a word 4-gram LM with whole-lattice re-scoring. We came up with in-domain n-gram LMs by training on the MARVEL training transcripts with modified Kneser-Ney smoothing in SRILM [49].

We developed two ASR systems by fine-tuning the pre-trained model on either MARVEL or JCCOCC-MoCA and pairing the resulting fine-tuned models with the n-gram word-level language model. We evaluated the performance of the systems on the participants' speech of the MARVEL test set and results are given in Table 4. Not surprisingly, the system which adopts in-domain MARVEL training data for ASR fine-tuning and LM training performed better.

[1]https://github.com/nryant/dscore

[2]https://k2-fsa.github.io/k2

Table 4: *ASR character error rates (CER%) on the participants' speech in the MARVEL test set*

Sys.	ASR fine-tuning data	LM training data	CER (%)
A	MARVEL	MARVEL	16.27
B	JCCOCC-MoCA	MARVEL	20.90

6. NCD Detection

Based on the speech transcriptions, the proposed pipeline system proceeds to distinguish NCD participants (i.e. the mild- and Major-NCD groups) from the cognitively healthy participants. We adopt a similar methodology as [24], which was previously applied to the English NCD detection dataset ADReSS20. The detection system consists of a pre-trained language model (PLM) text encoder and a back-end classifier. The PLM text encoders were fine-tuned offline on the MARVEL training data to adapt to the cognitive assessment texts. The resultant text embeddings were input into the NCD detection classifiers. Model snapshots during the fine-tuning at the final three update epochs were used to produce separate text embeddings, and their resultant NCD detection predictions were combined by majority voting to reduce the risk of over-fitting and to enhance robustness. This system achieved state-of-the-art AD detection accuracy of 87.50% with RoBERTa as the PLM and SVM as the classifier (Table 5, system 4).

We evaluated our experiments with either ground truth manual transcripts on the transcribed subset, which could provide a reliable baseline, or ASR transcripts generated by system A (Table 4) on the full set, which could enable fully automated NCD screening. The manual systems were trained and tested both with manual transcripts, while manual transcripts were still adopted as training data when evaluating the ASR transcripts system based on the MARVEL data fine-tuned ASR system A (Table 4), resulted in a hybrid transcripts system. The aim was to avoid the impact of data leakage from the ASR stage. PLM fine-tuning for each system was conducted with its corresponding classifier training data.

As an initial work on MARVEL, we adopted straightforward text pre-processing techniques. For each participant, the transcripts of only the participant side from six cognitive tests were concatenated as PLM inputs. Participant transcripts with length exceeding the PLM max input length (512 tokens) were cut into text segments by a sliding window and without overlaps. The participant-level decision was aggregated from segment-level decisions by majority voting. All non-Chinese-character annotations in manual transcripts were filtered out to maintain consistency with ASR transcripts.

In this pipeline system, RoBERTa[3] and multilayer perceptron (MLP) were selected as the PLM and classifier. The classifier type, as well as the hyper-parameters for PLM fine-tuning and MLP training, were selected by methods in [24] to optimize the detection performance. Out-of-vocabulary characters in the training data were added to the RoBERTa tokenizer before fine-tuning. The evaluation of each system is conducted based on 5 runs using different seeds for Pytorch random initialization[4]. The best and the average results are reported.

Table 5 shows the performances of the NCD detection systems. We observe the following main trends primarily based

[3]from https://huggingface.co/roberta-base, pre-trained with simplified Chinese

[4]https://pytorch.org/docs/stable/notes/randomness.html

Table 5: *NCD detection performance of the RoBERTa + MLP detection module with or without PLM fine-tuning, using manual transcripts (trans.) for classifier training, evaluated on manual trans. on the labeled (i.e. partial) test partition or ASR trans. on the full test set of the MARVEL dataset. The best and average results from 5 runs are reported for fine-tuned PLM systems. "Acc.", "Prec.", "Rec." and "F1" represent accuracy, precision, recall and F1 scores, NCD as the relevant class. Reference results of the pipeline system with the RoBERTa + SVM detection module on the ADReSS20 English dataset are also listed.*

Sys.	Dataset	Train Trans.	Test Trans.	Test Partition	Fine -tuning	Exp. Run	Test Result							
							Best				Average			
							Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
1	ADReSS	Manual	Manual	Full (48)	×	1	0.7083	0.7273	0.6667	0.6957	-			
2					√	5	0.8542	0.8339	0.7833	0.8058	0.8125	0.8696	0.8333	0.8511
3			ASR	Full (48)	×	1	0.7292	0.6667	0.9167	0.7719	-			
4					√	5	**0.8750**	0.8000	1.0000	0.8889	0.8208	0.8087	0.8417	0.8219
5	MARVEL	Manual	Manual	Labeled Participants (46)	×	1	0.7826	0.7692	0.8333	0.8000	-			
6					√	5	**0.8043**	0.7586	0.9167	0.8302	0.7522	0.7312	0.8333	0.7782
7			ASR Sys.A	Full (120)	×	1	0.7167	0.7083	0.3864	0.5000	-			
8					√	5	0.7167	0.6923	0.4091	0.5143	0.6900	0.6234	0.4091	0.4855
9				Labeled Participants (46)	×	1	0.6087	0.7500	0.3750	0.5000	-			
10					√	5	**0.6957**	0.7500	0.6250	0.6818	0.6174	0.7543	0.4000	0.5120

on the accuracy (Acc.) and recall (Rec., i.e. NCD sensitivity) scores. First, manual transcripts outperform the ASR transcripts (system 6 versus system 10), which may be attributed to a number of reasons, including ASR errors, or the mismatch between manual-based training data and ASR-based test data, affecting downstream NCD detection. Second, in contrast to its efficacy on English data, the strategy of fine-tuning PLM only generates performance gains in the best run (Table 5, systems 6&10, columns 8-11) and leads to degradation on average performances(Table 5, systems 6&8, column 12-15), indicating the need for further explorations on adapting PLM to cognitive assessment texts and improving NCD detection robustness.

7. Conclusions

This paper presents the development of an enhanced pipeline system for automated screening of neurocognitive disorders based on raw speech recordings from a two-way interaction between an assessor and participating in spoken Cantonese. The system was constructed by tandem components of the speaker diarization module for the participants' speech extraction, the ASR for speech transcription, the language model for text encoding, and the classifier to perform NCD detection from text embeddings. Compared with previous approaches, we made enhancements including special provision for overlapped speech detection in the speaker-turn aware speaker diarization, and in-domain data fine-tuning for both self-supervised pre-trained ASR and PLM. A competitive detection accuracy of 87.5% was achieved on the ADReSS20 corpus with a partial pipeline system with oracle diarization. Meanwhile, the full system produced 80.43% accuracy on the transcribed partition and 69.57% accuracy on the full test set on the Cantonese MARVEL corpus. Our future work includes spoken language feature selection and analysis of the effect of dysfluencies.

8. References

[1] J. Appell *et al.*, "A study of language functioning in Alzheimer patients," *Brain and language*, vol. 17, no. 1, pp. 73–91, 1982.

[2] J. L. Cummings *et al.*, "Alzheimer's disease and Parkinson's disease: comparison of speech and language alterations," *Neurology*, vol. 38, no. 5, pp. 680–680, 1988.

[3] K. J. Han *et al.*, "Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization," *IEEE Transactions on Audio, Speech, and Language Processing*, 2008.

[4] S. S. Xu *et al.*, "Age-invariant speaker embedding for diarization of cognitive assessments," in *ISCSLP*, 2021.

[5] Q. Lin *et al.*, "LSTM based similarity measurement with spectral clustering for speaker diarization," *arXiv preprint arXiv:1907.10393*, 2019.

[6] S. S. Xu *et al.*, "Speaker turn aware similarity scoring for diarization of speech-based cognitive assessments," in *Proc. APSIPA ASC*, 2021, pp. 1299–1304.

[7] K. C. Fraser *et al.*, "Linguistic features identify Alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[8] J. Weiner *et al.*, "Speech reveals future risk of developing dementia: Predictive dementia screening from biographic interviews," in *2019 IEEE ASRU*, 2019, pp. 674–681.

[9] S. Luz *et al.*, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge," *Proc. Interspeech 2020*, pp. 2172–2176, 2020.

[10] C. Frankenberg *et al.*, "Verbal fluency in normal aging and cognitive decline: Results of a longitudinal study," *Computer Speech & Language*, vol. 68, p. 101195, 2021.

[11] J. Li *et al.*, "A comparative study of acoustic and linguistic features classification for Alzheimer's disease detection," in *ICASSP*, 2021.

[12] Y. Fan *et al.*, "COMPARE: classification of morphological patterns using adaptive regional elements," *IEEE transactions on medical imaging*, 2006.

[13] F. Eyben *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, 2015.

[14] S. Petrov *et al.*, "A universal part-of-speech tagset," in *Proc. LREC*, 2012.

[15] C. K. Tomoeda *et al.*, "Speech rate and syntactic complexity effects on the auditory comprehension of Alzheimer patients," *Journal of Communication Disorders*, vol. 23, no. 2, pp. 151–161, 1990.

[16] J. Koo *et al.*, "Exploiting multi-modal features from pre-trained networks for Alzheimer's dementia recognition," in *INTERSPEECH*, 2020.

[17] A. Balagopalan *et al.*, "Comparing acoustic-based approaches for Alzheimer's disease detection," *arXiv preprint arXiv:2106.01555*, 2021.

[18] Z. S. Syed *et al.*, "Automated recognition of Alzheimer's dementia using bag-of-deep-features and model ensembling," *IEEE Access*, vol. 9, pp. 88 377–88 390, 2021.

[19] J. Li *et al.*, "Leveraging pretrained representations with task-related keywords for Alzheimer's disease detection," in *ICASSP 2023*, 2023.

[20] A. Balagopalan *et al.*, "To BERT or not to BERT: Comparing speech and language-based approaches for Alzheimer's Disease detection," *Proc. Interspeech*, 2020.

[21] J. Yuan *et al.*, "Disfluencies and fine-tuning pre-trained language models for detection of Alzheimer's disease." in *Proc. INTERSPEECH*, 2020.

[22] L. Ilias *et al.*, "Explainable identification of dementia from transcripts using transformer networks," *arXiv preprint arXiv:2109.06980*, 2021.

[23] M. Martinc *et al.*, "Temporal integration of text transcripts and acoustic features for Alzheimer's diagnosis based on spontaneous speech," *Frontiers in Aging Neuroscience*, p. 299, 2021.

[24] Y. Wang *et al.*, "Exploring linguistic feature and model combination for speech recognition based automatic ad detection," *INTERSPEECH*, 2022.

[25] L. Zhou *et al.*, "Speech Recognition in Alzheimer's Disease and in its Assessment," in *Proc. INTERSPEECH*, 2016.

[26] R. Pappagari *et al.*, "Automatic detection and assessment of Alzheimer disease using speech and language technologies in low-resource scenarios." in *Proc. INTERSPEECH*, 2021.

[27] H. O. MohamedShreif *et al.*, "Speech recognition for early detecting Alzheimer's disease by using machine learning algorithms," in *Proc. ICEMIS*, 2022.

[28] J. Becker *et al.*, "The natural history of Alzheimer's disease. description of study cohort and accuracy of diagnosis," *Arch. Neurol*, vol. 51, pp. 585–594, 1994.

[29] J. Galvin *et al.*, "The AD8," *Neurology*, vol. 65, pp. 559 – 564, 2005.

[30] C. Kan *et al.*, "The informant AD8 can discriminate patients with dementia from healthy control participants in an Asian older cohort." *Journal of the American Medical Directors Association*, vol. 20 6, pp. 775–779, 2019.

[31] A. Wong *et al.*, "The validity, reliability and clinical utility of the Hong Kong Montreal Cognitive Assessment in patients with cerebral small vessel disease," *Dementia and Geriatric Cognitive Disorders*, vol. 28, pp. 81 – 87, 2009.

[32] A. S. Chan, *Hong Kong List Learning Test*. Department of Psychological and Integrative Neuropsychological Rehabilitation Centre, The Chinese University of Hong Kong, 2006.

[33] D. Wechsler, *Wechsler Memory Scale - Fourth Edition*. The Psychological Corporation, 2009.

[34] R. Cheung *et al.*, "Confrontation naming in Chinese patients with left, right or bilateral brain damage," *Journal of the International Neuropsychological Society*, 2004.

[35] A. He *et al.*, "Developing an alternative to the frog story: Evidence from Cantonese/Mandarin-English bilinguals," in preparation.

[36] X. Gong *et al.*, "The Hong Kong Grocery Shopping Dialog Task (HK-GSDT): A quick screening test for neurocognitive disorders," *International Journal of Environmental Research and Public Health*, vol. 19, 2022.

[37] S.-H. Gao *et al.*, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 652–662, 2019.

[38] A. Nagrani *et al.*, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[39] J.-S. Chung *et al.*, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[40] Y. Fan *et al.*, "CN-Celeb: a challenging Chinese speaker recognition dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 7604–7608.

[41] L. Li *et al.*, "CN-Celeb: multi-genre speaker recognition," *Speech Communication*, vol. 137, pp. 77–91, 2022.

[42] M. Diez *et al.*, "Analysis of speaker diarization based on bayesian HMM with eigenvoice priors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 355–368, 2019.

[43] H. Bredin *et al.*, "Pyannote. audio: neural building blocks for speaker diarization," in *Proc. ICASSP*, 2020.

[44] F. Yu *et al.*, "M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge," in *Proc. ICASSP*, 2022.

[45] Y. Fu *et al.*, "Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario," *arXiv preprint arXiv:2104.03603*, 2021.

[46] A. Baevski *et al.*, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, H. Larochelle *et al.*, Eds., 2020.

[47] A. Babu *et al.*, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proc. INTERSPEECH*, 2022.

[48] M. Ott *et al.*, "FAIRSEQ: A fast, extensible toolkit for sequence modeling," in *Proc. NAACL-HLT*, 2019.

[49] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *Proc. ICSLP*, 2002.