HK-GenSpeech: A Generative AI Scene Creation Framework for Speech Based Cognitive Assessment

Vi Jun Sean Yong[1], *Serkan Kumyol*[1], *Pau Le Lisa Low*[2], *Suk Wai Winnie Leung*[1], *Tristan Braud*[1]

[1]Division of Integrative Systems and Design, The Hong Kong University of Science and Technology, Hong Kong SAR

[2]S.K. Yee School of Health Sciences, Saint Francis University, Hong Kong SAR

vjsyong@connect.ust.hk, skumyol@connect.ust.hk, lisalow@sfu.edu.hk, eewswleung@ust.hk, braudt@ust.hk

Abstract

Current methods of automated speech-based cognitive assessment often rely on fixed-picture descriptions in major languages, limiting repeatability, engagement, and locality. This paper introduces HK-GenSpeech (HKGS), a framework using generative AI to create pictures that present similar features to those used in cognitive assessment, augmented with descriptors reflecting the local context. We demonstrate HKGS through a dataset of 423 Cantonese speech samples collected in Hong Kong from 141 participants, with HK-MoCA scores ranging from 11 to 30. Each participant described the cookie theft picture, an HKGS fixed image, and an HKGS dynamic image. Regression experiments show comparable accuracy for all image types, indicating HKGS' adequacy in generating relevant assessment images. Lexical analysis further suggests that HKGS images elicit richer speech. By mitigating learning effects and improving engagement, HKGS supports broader data collection, particularly in low-resource settings.

Index Terms: Dementia detection, Speech-based cognitive assessment, Generative AI for healthcare

1. Introduction

Dementia is a progressive neurodegenerative disorder affecting an ageing population. While incurable, early diagnosis enables interventions that can improve patients' quality of life. Linguistic and acoustic biomarkers provide a non-invasive, cost-effective means of detection [1]. Initiatives like the ADReSS and ADReSSo challenges provide standardised datasets and tasks, using audio from the Boston Diagnostic Aphasia Examination Cookie Theft Picture (CTP) description task [2, 3, 4]. More recently, the ADRESS-M and TAUKADIAL challenge [5, 6] expanded efforts to multilingual approaches, emphasising linguistic diversity with other picture description tasks. However, most datasets still rely on fixed scenes, potentially limiting speech variability. Besides, under-represented languages, such as Cantonese, remain largely unaddressed, despite advancements in multilingual and trans-lingual methods [7, 5]. Generative AI, though under-explored in dementia detection, presents an opportunity to diversify stimuli and enhance engagement by tailoring assessments that relate to local populations [8, 9].

This work introduces **HK-GenSpeech (HKGS)**, a generative AI-based picture description framework for dementia detection that extends speech-based assessment beyond the CTP. The proposed framework employs a multi-stage pipeline to generate culturally adaptable and task-relevant scenes that elicit speech responses for predicting continuous HK-MoCA [10] scores. Compared to datasets like ADReSS (156 samples)[2] and ADReSSo (237 samples)[3], which rely on the CTP, HKGS uses dynamically generated scenes. This approach prevents memorisation in repeated tests and can adapt to the cultural and linguistic context of Cantonese-speaking participants in Hong Kong. Notably, it enables using multiple scenes per participant to increase the diversity of the dataset, which may be beneficial when recruiting additional participants poses a challenge.

To validate this approach, we used HKGS to collect a dataset of Cantonese speech samples from 141 community-dwelling participants in Hong Kong, aged 55–94, with HK-MoCA scores ranging from 11 to 30. Each participant provided three samples: (1) CTP description, (2) a fixed HKGS AI-generated picture, and (3) a dynamic HKGS AI-generated picture. All three conditions yielded similar results, with MAE values between 3.84 and 4.14, demonstrating the adequacy of HKGS AI-generated images in eliciting meaningful speech for cognitive assessment. Predictions remained consistent across conditions, indicating improved repeatability. A model incorporating all three samples per participant marginally outperformed individual models, suggesting that HKGS can augment data when participant numbers are limited.

2. The HKGS Framework

The HKGS framework follows three primary goals:

1. **Continuous Score Prediction:** The framework targets predicting continuous HK-MoCA scores rather than binary classification, for finer distinctions in cognitive performance, especially in Mild Cognitive Impairment (MCI) and dementia.
2. **Improved Test Repeatability:** Dynamically generated scenes mitigate the risk of learning effect and reduce boredom, improving repeatability for longitudinal studies.
3. **Cultural Relevance:** Generating scenes that relate to the local culture to better reflect the participants' environment, improving comfort, engagement, and inclusivity.

HKGS relies on an image generation pipeline that converts textual and visual cues into AI-generated scenes as shown in Figure 1. The pipeline first distils the meaningful features of the CTP into guidelines, which are used to generate unique images.

2.1. Image Feature Extraction and Guideline Derivation

The CTP is efficient in cognitive assessment by stimulating the visuospatial capabilities of the subject through a detailed scene that elicits narrative descriptions in a familiar setting. HKGS replicates such characteristics in AI-generated pictures. First, we use GPT-4o [11] vision capabilities to generate text descriptions that capture the central events, background context, and relational dynamics of the CTP. We then refine the generated description by passing Cummings' [12] paper describing regions of interest in the CTP for speech-based assessments, along with the text descriptors to further distil the CTP into a set of im-

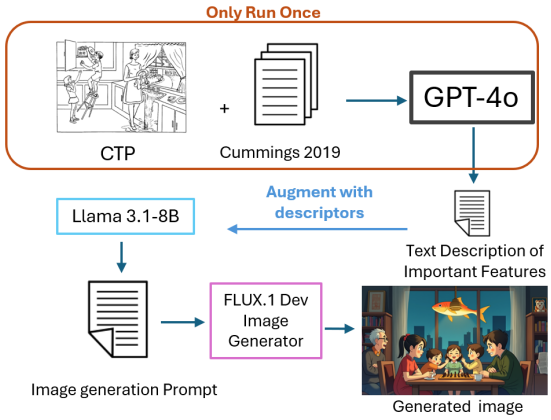

Figure 1: *Image generation pipeline. The core features of the CTP are first extracted and augmented with culturally relevant descriptors to generate prompts for creating AI-generated images.*

age design principles. Finally, we instructed GPT4o to convert these guidelines into a guiding prompt for a secondary LLM to create the image generation prompts.

2.2. Design Principles for Diagnostic Image Generation

From the previous step, we extract a set of distilled design principles mainly based on Cumming's work[12]:

- D1 **Salience of Information** – The image should feature a central, easily recognisable event, with background elements providing context without distraction.
- D2 **Semantic Categories** – A variety of animate and inanimate entities should be included, allowing for descriptions that use both general and specific terms.
- D3 **Referential Cohesion** – Multiple characters and objects should be present, requiring clear referential identification through pronouns and anaphoric references.
- D4 **Causal and Temporal Relations** – The scene should imply logical event sequences and causal relationships.
- D5 **Mental State Language** – Characters should exhibit emotions or intentions requiring inferential language.
- D6 **Structural Language and Speech** – The composition should encourage diverse syntactic structures, supporting both simple and complex sentences.
- D7 **General Cognition and Perception** – The layout should be balanced while promoting attention to detail.

HKGS adheres to these guidelines to generate images that elicit meaningful responses for cognitive assessment. Fig. 3 presents the control image generated using these design principles. It features a family playing chess (D1), with diverse animate and inanimate entities, including family members and household objects (D2, D3). Characters' interactions imply a logical sequence of actions, and various relationships (D3,D4), requiring complex sentences to describe (D6), while their facial expressions and body language convey diverse emotions and intentions, including pride, surprise, and worry (D5). The scene presents a broad story while featuring many small details, such as items on the bookshelves and characters out of the view (D7).

2.3. The Image Generation Pipeline

HKGS uses the guidelines in subsection 2.2 in a multi-stage image generation pipeline as follows:

Table 1: *Summary Statistics Speech-Based Cognitive Assessment Datasets*

Dataset	Lang.	Pic. Task	# Samples/Part.	Cogn. Test
HK-GenSpeech	Cantonese	CTP (Baseline), GC, GU	423/141	HK-MoCA
ADReSS[2]	English	CTP	156/156	MMSE
ADReSSo[3]	English	CTP + Fluency Task	237/237	MMSE + AD
TAUKA-DIAL[16]	English, Mandarin	3 Pic. (Eng.) 3 Pic. (Mand.)	507/169	MoCA + MCI
ADReSS-M[6]	English, Greek	CTP (English), 1 Pic. (Greek)	225/225 (English) 46/46 (Greek)	MMSE + AD

Figure 2: *Sample pictures from the HKGS accompanying dataset*

1. **Augmenting Prompt Descriptors** The guiding prompt created by GPT-4o in subsection 2.1 is first augmented with culturally and task-relevant modifiers (e.g., "Generate an image generation prompt in a single paragraph format, do not explain anything", "set in Asia", "animation style") to adapt the scene to the local population.
2. **Refining Scene Prompts:** A Llama 3.1-8B model [13, 14], chosen for its strong performance given its size and ability to be run locally at minimal cost, takes in the augmented prompt and generates dynamic prompts tailored to image generation tools. Paired with a random seed, this step ensures that the generated scene descriptions are unique with every run, yet remain consistent in appearance and adherence to the guidelines in subsection 2.2.
3. **Image Synthesis:** The final prompts are fed into a workflow using Flux.1 Dev [15] to create the scene images. Each image is then assigned a unique ID to facilitate tracking and association to the recipient participant during analysis.

Fig. 2 shows some of the scenes generated by this pipeline.

2.4. Data Collection, Dataset and Task Design

We use the images generated in subsection 2.3 to collect a dataset of Cantonese-language image descriptions. The study involved 141 community-dwelling, Cantonese-speaking participants recruited from seven common gathering places for older

Figure 3: *Gen-Control(GC) Task - Example of an image created using the HKGS design principles*

adults in Hong Kong (community centres, day centres, and churches) via convenience sampling. Differences in cognitive function were observed across sites, with church participants presenting higher HK-MoCA scores, possibly due to the degree of autonomy and cognition required to participate in such activities [17, 18]. Participants were aged 55 to 94 (75.73 ± 8.67), 38 males and 103 females, with self-reported education levels ranging from no formal schooling[1] to master's degree (6.39 ± 3.89 years). Their assessed HK-MoCA scores range between 11 and 30 (21.1 ± 5.53). Participants were required to have good eyesight with or without correction, be able to speak fluent Cantonese, and be able to hear and understand spoken instructions. The total duration of all samples is 7h 38m with an average length of 1m 56s.

The interviews were conducted indoors in a one-to-one setting. Participants were first informed of the study's procedure, goals, and outcomes before providing written consent. They then completed the HK-MoCA test, administered by a pool of researchers. Following the test, participants moved to a separate room for speech description recordings after a brief intermission. During this session, they viewed and verbally described three images (all images were manually screened for major anatomical errors, and fewer than 8% required replacement) in a predetermined order to ensure dataset consistency.

1. **Baseline (CTP):** The original, static Cookie Theft Picture serving as a reference.
2. **Gen-Control (GC):** A standardised AI-generated image produced using the framework for control testing. (See Fig. 3)
3. **Gen-Unique (GU):** A unique, dynamically generated image assigned randomly to each participant.

Participants were asked to "tell the researcher everything they saw in the picture". The three pictures were presented sequentially, mirroring the protocols of the ADReSS and ADReSSo datasets [2, 3]. The protocol was approved by the university's IRB and we followed typical guidelines for gathering consent and conducting design activities with people with MCI or dementia [19, 20].

3. Evaluation

This section evaluates the application of HKGS in speech-based cognitive assessments by comparing models trained on the CTP, GC, and GU datasets.

[1] 13% of interviewed participants did not receive any form of education, due to historical factors in the region in the mid-20th century.

Figure 4: *Wav2Vec2 HK-MoCA Predictor Model Diagram*

Table 2: *Performance Metrics for Speech Assessment Models*

Model	MAE	RMSE	R^2	Pearson_R	Bias	% Within 1-pt
Wav2Vec2 + Neural Network Regressor						
CTP	3.87	**4.63**	**0.30**	0.55	0.57	12.77%
GC	4.14	5.19	0.11	0.47	1.83	16.43%
GU	3.84	4.76	0.25	0.53	**0.56**	14.89%
Combined	**3.64**	4.67	0.28	**0.57**	1.03	**19.87%**
eGeMAPs V2 + Support Vector Regressor						
CTP	4.56	5.30	0.04	0.23	**0.66**	8.25%
GC	4.29	5.13	0.07	0.31	0.86	10.90%
GU	4.29	5.22	0.07	0.32	0.94	**12.73%**
Combined	**4.15**	**4.95**	**0.16**	**0.42**	0.69	11.23%

3.1. Experiment Setup

3.1.1. Dataset Preparation and Audio Preprocessing

The CTP, GC, and GU datasets each contain an equal number of samples—one per participant per set—to ensure a consistent evaluation across tasks. Stratified five-fold cross-validation (CV) was applied, with training and validation splits generated using the same seed (42) to maintain consistency.

For each sample, we first filter the background noise with DeepFilterNet [21]. The samples were then imported into Audacity [22] to truncate silent regions to a maximum of 0.75 seconds, as some samples contained extended periods of silence [23]. Next, samples were labelled as *Interviewer*, *Participant*, or *Irrelevant* – *Irrelevant* being used as a catch-all category for sections unrelated to the picture description task. Regions labelled as *Interviewer* or *Irrelevant* were removed. Samples were normalised in amplitude and exported as 16-bit mono-channel .wav files with a sample rate of 16 KHz. Each audio sample was then divided into 15-second chunks. Any chunks shorter than 10 seconds were discarded.

3.1.2. Model Architecture and Training Setup

We developed a regression model using a fine-tuned Cantonese Wav2Vec2 model [21, 24] as the backbone. The model extracts embeddings from the preprocessed audio from Section 3.1, which are then mean-pooled and fed into a dense network. The dense network consists of two layers with 64 neurons each, ReLU activation, and a dropout rate of 0.3. The final output layer comprises a single neuron. The model was trained to minimise the mean squared error (MSE) loss (see Fig. 4).

Four models were trained: one for each task-specific dataset (CTP, GC, GU) and one using all available data, stratified by participant. The models were trained with a learning rate of 5×10^{-4} and a batch size of 8. Each model was trained up to 20 epochs per CV split, with early stopping applied after three epochs of patience, using "Eval MAE" to determine the best-performing model. For comparison, we also trained baseline regression models using eGeMAPS V2 features in combination with an SVR, as established in previous work [2, 3].

3.2. Experiment Results

3.2.1. Prediction Performance

From Table 2 baseline models using eGeMAPSv2 features with an SVR are outperformed by our approach, confirming its ef-

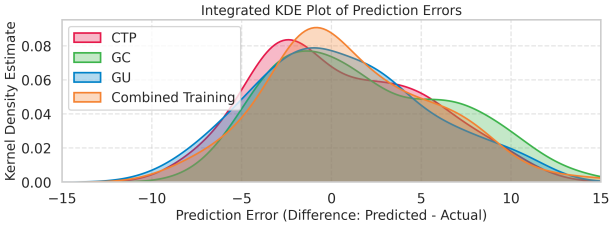

Figure 5: *KDE Plot Comparing Model Prediction Errors. The combined training approach results in the highest proportion of predictions within 1 point of the ground truth.*

Table 3: *Summary of Linguistic Metrics*

Model	TTR	POS Div.	Con. WR	Lex. Den.	Avg. UL
CTP	0.135	0.648	0.480	0.670	5.99
GC	0.142	0.654	0.480	0.640	6.09
GU	0.165	0.657	0.470	0.660	5.91

TTR: Type-Token Ratio. **POS Div**.: Part-of-Speech Diversity. **Con. WR**: Content Word Ratio. **Lex. Den**.: Lexical Density. **Avg. UL**: Average Utterance Length.

fectiveness. Wav2Vec2 models trained on AI-generated images perform comparably to those trained on the CTP for predicting HK-MoCA scores. A naive baseline always predicting the mean (21.1) yields an MAE of 4.57 and an RMSE of 5.51. By contrast, the GU model achieves an MAE of 3.84 and an RMSE of 4.76—close to the CTP (MAE = 3.87, RMSE = 4.63)—whereas the GC model produces higher errors (MAE = 4.14, RMSE = 5.19), possibly due to greater variability with static images.

Repeated measures ANOVA indicates a significant effect of the split/task condition ($F(2,278) = 18.24$, $p < 0.001$). However, Tukey post hoc comparisons revealed no significant difference between CTP and GU (mean difference = 0.00, $p = 1.000$). Both differ significantly from GC (mean difference = 1.24, $p = 0.0027$). The combined dataset achieves the lowest MAE (3.64), a competitive RMSE (4.67), and the highest Pearson's R (0.57) and R^2 (0.28), predicting scores within one point in nearly 20% of cases (under 13% for CTP alone). These findings suggest that broader linguistic diversity can enhance generalisation without compromising performance.

3.2.2. Linguistic Analysis

Performance differences across tasks may stem from subtle linguistic variations in elicited speech. As shown in Table 3, GU responses have a higher TTR (0.165) and slightly greater POS diversity (0.657) than CTP (0.135 and 0.648, respectively), suggesting more varied lexical usage and additional cues for predicting cognitive performance. In contrast, CTP responses exhibit a slightly higher content word ratio (0.480) and lexical density (0.670), suggesting more focused content.

Average utterance length is consistent across conditions (6 words), suggesting that qualitative rather than quantitative differences in speech drive the observed effects. Although engagement was not directly measured, linguistic differences and noun usage patterns (e.g., "thing" (東西) in CTP vs. "child" (孩子) in GC and GU) suggest greater spontaneity in the latter. These modest lexical diversity differences may contribute to the model's predictive performance.

3.2.3. Bias and Accuracy Metrics

According to Table 2, bias analysis shows relatively low mean bias error (MBE) across models. The GU model (MBE 0.56) aligns closely with the CTP (MBE 0.57), whereas GC exhibits a higher bias (MBE 1.83), possibly due to the systematic influence of the static AI-generated image. The combined model, with an MBE of 1.03, remains within acceptable range and achieves the highest percentage of predictions within 1 point of the ground truth (19.87%), followed by GC (16.43%), GU (14.89%) and CTP (12.77%). This suggests that combined training effectively balances bias reduction and precision, benefiting from exposure to a broader range of linguistic patterns.

3.3. Discussion

We summarise the main findings and opportunities of HKGS, and discuss some limitations.

Adequacy for cognitive assessment: Images generated with HKGS yield HK-MoCA prediction scores comparable to the reference CTP descriptions. For most subjects, average pairwise differences remain below 2.5 points, demonstrating that HKGS adequately replicates the CTP features that elicit speech suitable for cognitive assessment.

Engagement and cultural relatability: During data collection, several participants misinterpreted the CTP, especially the scene outside the window, due to its rarity in Hong Kong. The adult woman was often seen as a domestic worker, reflecting local societal contexts. Participants were more engaged with the GC and GU scenes, leading to better scene recognition and more spontaneous descriptions. Linguistic analysis supports this observation, showing greater lexical variety in AI scenes.

Repeatability and Data Diversity: With dynamic image generation, HKGS prevents participants from memorising their responses, mitigating learning effects. This enables the collection of multiple samples over a single session. Training on all samples improves performance, suggesting that the prediction benefits from the increased lexical diversity and that the approach can function as data augmentation in low-resource settings. As such, it opens new opportunities from collecting data on underrepresented populations to conducting longitudinal studies.

Data bias: While the generated images can adapt to local contexts, their effectiveness is limited by the diversity of the training data of image generation models. Certain cultural or demographic contexts are typically under-represented, leading the generated scenes not to authentically reflect those experiences. For instance, we noted a low number of variations among characters' appearances when representing East Asian populations.

4. Conclusion

This work introduces the HKGS framework, enabling the creation of AI-based images for speech-based cognitive assessment. The framework allows the creation on-demand of images that emulate the features of the CTP while adapting to the local cultural context to elicit richer speech. Experiments show that AI-generated images perform comparably to the Cookie Theft Picture (CTP) in predicting HK-MoCA scores while providing greater linguistic diversity. Prediction results also emphasize the repeatability of the experience, with sequences of images generated with HKGS yielding similar accuracy.

HKGS creates new opportunities for facilitating data collection for speech-based cognitive assessment in environments where participants are scarce, in low-resource languages and locations, and opens the possibility of more longitudinal studies.

5. Acknowledgements

1. The work described in this paper was supported by the HKUST Center for Aging Science (project ID: Z1010)
2. We want to thank Haven of Hope Christian Services Hong Kong for their support in the recruitment of study participants.
3. We want to thank International Christian Assemblies Hong Kong for their support in the recruitment of study participants.

6. References

[1] National Institute on Aging, "Alzheimer's disease fact sheet," https://www.nia.nih.gov/health/alzheimers-and-dementia/alzheimers-disease-fact-sheet, n.d., accessed: 2025-01-17.

[2] S. Luz, F. Haider, S. de la Fuente, D. Fromm *et al.*, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," *Proc. Interspeech*, pp. 2172–2176, 2020.

[3] S. Luz *et al.*, "Detecting cognitive decline using speech only: The adresso challenge," *Proc. Interspeech*, pp. 3942–3946, 2021.

[4] H. Goodglass, E. Kaplan, and B. Barresi, *Boston Diagnostic Aphasia Examination–Third Edition (BDAE-3)*. Philadelphia, PA: Lippincott Williams & Wilkins, 2001.

[5] S. Luz *et al.*, "Connected speech-based cognitive assessment in chinese and english," *Proc. Interspeech*, pp. 1123–1127, 2024.

[6] S. Luz, F. Haider, D. Fromm, I. Lazarou, I. Kompatsiaris, and B. MacWhinney, "An Overview of the ADReSS-M Signal Processing Grand Challenge on Multilingual Alzheimer's Dementia Recognition Through Spontaneous Speech," *Proc. Interspeech*, vol. 5, pp. 738–749, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10474114

[7] M. Hoang *et al.*, "Translingual language markers for cognitive assessment from spontaneous speech," *Proc. Interspeech*, pp. 1024–1030, 2024.

[8] T.-B. Chen, C.-Y. Lin, K.-N. Lin, Y.-C. Yeh, W.-T. Chen, K.-S. Wang, and P.-N. Wang, "Culture qualitatively but not quantitatively influences performance in the boston naming test in a chinese-speaking population," *Dementia and Geriatric Cognitive Disorders Extra*, vol. 4, no. 1, pp. 86–94, 2014. [Online]. Available: https://doi.org/10.1159/000360695

[9] Z. Zhang, L. Cui, L. Huang, Y.-H. Guan, F. Xie, and Q.-H. Guo, "Development and validation of the chinese naming test (cnt): Diagnostic efficacy and correlation with alzheimer's disease biomarkers," *Journal of Alzheimer's Disease*, vol. 104, no. 4, pp. 1259–1269, 2025. [Online]. Available: https://doi.org/10.1177/13872877251324100

[10] A. Wong, Y. Y. Xiong, P. W. Kwan, A. Y. Chan, W. W. Lam, K. Wang, W. C. Chu, D. L. Nyenhuis, Z. Nasreddine, L. K. Wong, and V. C. Mok, "The Validity, Reliability and Clinical Utility of the Hong Kong Montreal Cognitive Assessment (HK-MoCA) in Patients with Cerebral Small Vessel Disease," vol. 28, no. 1, pp. 81–87. [Online]. Available: https://karger.com/DEM/article/doi/10.1159/000232589

[11] OpenAI, "Introducing gpt-4o," https://openai.com/index/hello-gpt-4o/, 2025, accessed: 2025-01-20.

[12] L. Cummings, "Describing the cookie theft picture: Sources of breakdown in alzheimer's dementia," *Pragmatics & Society*, vol. 10, no. 2, pp. 151–174, 2019.

[13] M. AI, "Llama 3.1: Open and efficient foundation language models," https://github.com/meta-llama/llama-models, accessed: January 17, 2025.

[14] A. Dubey *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024, accessed: January 17, 2025. [Online]. Available: https://arxiv.org/abs/2407.21783

[15] Black Forest Labs, "FLUX.1 Dev: Frontier AI Image Generator Model," https://flux1ai.com/dev, accessed: January 17, 2025.

[16] G. Gosztolya and L. Tóth, "Combining acoustic feature sets for detecting mild cognitive impairment in the taukadial challenge," *Proc. Interspeech*, pp. 4356–4360, 2024.

[17] I. S. Nelson, K. Kezios, M. Elbejjani, P. Lu, K. Yaffe, and A. Zeki Al Hazzouri, "Associations of Religious Service Attendance With Cognitive Function in Midlife: Findings From The CARDIA Study," vol. 78, no. 4, pp. 684–694. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10413813/

[18] T. D. Hill, A. M. Burdette, J. L. Angel, and R. J. Angel, "Religious Attendance and Cognitive Functioning Among Older Mexican Americans," vol. 61, no. 1, pp. P3–P9. [Online]. Available: https://doi.org/10.1093/geronb/61.1.P3

[19] S. Suijkerbuijk, H. H. Nap, W. A. Ijsselsteijn, M. M. Minkman, and Y. A. de Kort, "'i already forgot half of it'–interviewing people with dementia for co-designing an intelligent system," *Human–Computer Interaction*, vol. 39, no. 3-4, pp. 225–256, 2024.

[20] S. Y. Kim, "The ethics of informed consent in alzheimer disease research," *Nature Reviews Neurology*, vol. 7, no. 7, pp. 410–414, 2011.

[21] H. Schröter, T. Rosenkranz, A. N. Escalante-B., and A. Maier, "DeepFilterNet: Perceptually motivated real-time speech enhancement," in *INTERSPEECH*, 2023.

[22] Muse Group and contributors, "Audacity: Free audio editor and recorder," 2025, version 3.3.3, licensed under the GNU General Public License, Version 3, accessed January 17, 2025. [Online]. Available: https://www.audacityteam.org/

[23] M. R. Kumar *et al.*, "Dementia detection from speech using machine learning and deep learning architectures," *Sensors*, vol. 22, no. 23, p. 9311, Nov 2022.

[24] H. Lovenia *et al.*, "Ascend: A spontaneous chinese-english dataset for code-switching in multi-turn conversation," *Proc. ACL*, pp. 1234–1240, 2022.