TalkTive: A Conversational Agent Using Backchannels to Engage Older Adults in Neurocognitive Disorders Screening

Zijian Ding
College of Information Studies, University of Maryland College Park
USA

Jiawen Kang
Department of SEEM, The Chinese University of Hong Kong
HKSAR

Tinky Oi Ting HO
Department of Psychology, The Chinese University of Hong Kong
HKSAR

Ka Ho Wong
Department of SEEM, The Chinese University of Hong Kong
HKSAR

Helene H. Fung
Department of Psychology, The Chinese University of Hong Kong
HKSAR

Helen Meng
Department of SEEM, The Chinese University of Hong Kong
HKSAR

Xiaojuan Ma
Department of CSE, Hong Kong University of Science and Technology
HKSAR

ABSTRACT

Conversational agents (CAs) have the great potential in mitigating the clinicians' burden in screening for neurocognitive disorders among older adults. It is important, therefore, to develop CAs that can be engaging, to elicit conversational speech input from older adult participants for supporting assessment of cognitive abilities. As an initial step, this paper presents research in developing the backchanneling ability in CAs in the form of a verbal response to engage the speaker. We analyzed 246 conversations of cognitive assessments between older adults and human assessors, and derived the categories of reactive backchannels (e.g. "hmm") and proactive backchannels (e.g. "please keep going"). This is used in the development of *TalkTive*, a CA which can predict both timing and form of backchanneling during cognitive assessments. The study then invited 36 older adult participants to evaluate the backchanneling feature. Results show that proactive backchanneling is more appreciated by participants than reactive backchanneling.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **Auditory feedback**; *Empirical studies in interaction design*; **Sound-based input / output**.

KEYWORDS

Backchanneling, Conversational Agents, Older Adults

ACM Reference Format:
Zijian Ding, Jiawen Kang, Tinky Oi Ting HO, Ka Ho Wong, Helene H. Fung, Helen Meng, and Xiaojuan Ma. 2022. TalkTive: A Conversational Agent Using Backchannels to Engage Older Adults in Neurocognitive Disorders Screening. In *CHI '22: CHI Conference on Human Factors in Computing Systems, April 30–May 06, 2022, New Orleans, LA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3491102.3502005

CHI '22, April 30–May 06, 2022, New Orleans, LA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3502005

1 INTRODUCTION

The rapidly ageing global population is imposing challenges to healthcare systems across nations [64]. Neurocognitive disorders (NCD), such as dementia, are particularly common in older adults [4]. The global cost of NCD exceeded the threshold of US$1 trillion in 2018 [70]. In Hong Kong, the estimated cost of institutional and informal care for older adults with NCD in 2036 is HK$31.2 billion (US$4 billion) [104]. Besides leading to an enormous financial burden on the society, NCD has a negative impact on the quality of life of older adults, their families, and caretakers, as well as on the workload of clinicians [70]. To look on the bright side, the negative symptoms associated with NCD have been shown to be controllable if patients can get access to early diagnoses and timely preventive interventions [2, 21]. This presents a necessity for a scalable approach in screening cognitive impairments among older adults. Current NCD screening and diagnosis tests, such as the Montreal Cognitive Assessment (MoCA) [60], are mainly conducted as in-person tests by clinical professionals [89, 100]. In-person assessments face limitations due to various factors, such as limited accessibility to the tests due to the lower mobility of some older adults, and shortages and inter-rater variabilities of clinicians [50].

To overcome these limitations, researchers are working on alternative solutions for NCD screening. Since NCDs are often manifested as communicative impairments, machine learning (ML) algorithms are offering a new type of support for NCD screening [3, 13, 22, 38, 47, 55, 56, 66, 72, 97, 100]. Therefore, a potential solution is to integrate speech analytics into a conversational agent (CA) to support highly accessible voice-based web applications that can interact with older adults through spoken language for screening NCD. Older adults can easily access web applications, which can collect their conversational speech data automatically, inexpensively and longitudinally. This opens up the possibility of

detecting subtle changes in an individual's cognitive abilities over time, to enable early detection of cognitive decline. Such interactive web applications may be able to engage older adults throughout the process of cognitive assessments, e.g. in responding to a series of questions from the CA, or performing a set of requested tasks. In particular, older adults experiencing cognitive decline may need special responses from their interlocutors, such as "elderspeak" [36]. To fulfill this requirement of communicating with older adults effectively, we may reference the inter-personal communication strategies adopted by expert assessors who are well-trained health professionals in engaging participants as they conduct cognitive assessments. Backchanneling is one of such strategies that is proven to be effective [35, 99]. Backchanneling refers to verbal responses such as "uh-huh" or "hmm", non-verbal responses such as nodding, smiling and gesturing, as well as *both* verbal and non-verbal responses simultaneously given by the listener when his/her counterpart is speaking [7, 9, 12, 73]. As a form of basic human interaction [29], backchanneling is displayed to show the engagement of the listener and to encourage the speaker to continue speaking without interrupting him/her. The importance of backchanneling has been stated in previous works such as showing understanding and engagement to the speakers [6, 30, 73], and establishing empathy between speakers and listeners [78].

In addition, previous work have also discussed the proactive characteristics of backchanneling. It has been shown that comprehension and production can co-exist in conversations [15]. Instead of only being passive recipients of information with reactive backchanneling, listeners can play an active role in a conversation as co-narrators [6]. In other words, listeners can give proactive backchanneling such as "please keep going". Ortega et al. [65] briefly discussed how to build a proactive listening system using backchannels to influence speakers, which is the first to model backchanneling from a proactive perspective. However, concrete definitions and methodologies of conducting proactive backchannels specific to the conversational context are largely missing in previous studies. To fill in this gap, we define two kinds of backchannels, reactive backchannels (RBCs) and proactive backchannels (PBCs), based on the theories of reactive and proactive backchanneling [65]:

- Reactive backchannels (RBCs): The listener responds to the previous speaker's utterance directly to show agreement or understanding without intending to take the floor. The response is generic (context-independent) and optional, thereby poses minimal interruption, i.e. the speaker does not need to wait for the response as he/she continues to speak. The response may serve as acknowledgement or assessment. Most RBCs are non-lexical; examples are "hmm", "oh", "yeah".
- Proactive backchannels (PBCs): The listener responds to the speaker to encourage the speaker to continue speaking without the intent to take the turn. This kind of response is also optional. The response might have more lexical words compared with RBC. Examples are "please keep going", "anything else?".

Based on the definitions of RBCs and PBCs, we aim to investigate their roles in engaging older adults in conversations with CAs to complete NCD screening tasks. Few previous studies have discussed task-based backchanneling, or potential opportunities in adaptive backchanneling depending on tasks and participants. To the best of our knowledge, this study may be one of the earliest to study backchanneling for older adult participants in a cognitive assessment context. Besides, the backchanneling behaviors have been proven to be language-dependent and culture-related [14, 17, 29]. To explore the potential for studying backchanneling in low-resource languages, we developed technology to support Cantonese speakers. Cantonese is a predominant Chinese dialect used in Hong Kong, and spoken by over 80 million native speakers worldwide [20]. To the best of our knowledge, this is also one of the first efforts that studies backchanneling in Cantonese. This prompts us to ask two research questions (RQs):

- **RQ1**: When do reactive and proactive backchannels happen in a task-driven conversation with older adults in Cantonese?
- **RQ2**: How do reactive and proactive backchannels provided by a task-oriented CA affect its conversation with older adults in Cantonese?

To answer these RQs, we first derived empirical patterns of how expert assessors provide RBCs and PBCs in real-world task-oriented NCD screening conversations, by analyzing 246 audio recordings of MoCA test conversations between older adult participants and assessors. Based on the results of this analysis, we developed data-driven ML models to predict proper timing for NCD screening CAs to deliver RBCs and PBCs, trying to mimic human strategies. Next, we built a proof-of-concept, backchanneling-enabled CA system called *TalkTive*, and conducted a between-subjects user study to evaluate older adults' perception of our system in MoCA test conversations, in comparison to a baseline system without backchanneling. A timeline of research activities presented in this paper is shown in Figure 1.

Our main contributions include: 1) developing a backchanneling algorithm to provide RBCs and PBCs with high performance, 2) generating automatic backchannels that offers a positive user experience for older adults, and 3) identifying the type of backchanneling (reactive versus proactive) and discovering that older adults are more receptive to RBCs than PBCs.

2 RELATED WORK

2.1 Conversational Agents in Healthcare Support for Older Adults

In face of an ageing global populations, it becomes increasingly important to develop technologies that offer scalable support for older adults, as well as alleviate the pressures of caregiving in the society. Related research has also attracted growing interests in the HCI community, e.g. [16, 27, 45, 48, 57, 94]. One of the main focuses of these HCI studies is to explore potential technological solutions that may reduce the expensive medical costs and enable older adults to manage their own care with greater independence [94], such as providing accessible health monitoring and disease screening solutions [18, 19, 39, 44, 46, 68]. Among various emerging and evolving technologies, conversational agents (CAs), defined as "systems that mimic human language and behavior to implement certain tasks for the user via a chat interface, either text-based or voice-based" [1], has unique advantages in healthcare surveys and disease screening tasks. Recent research in the adoption and

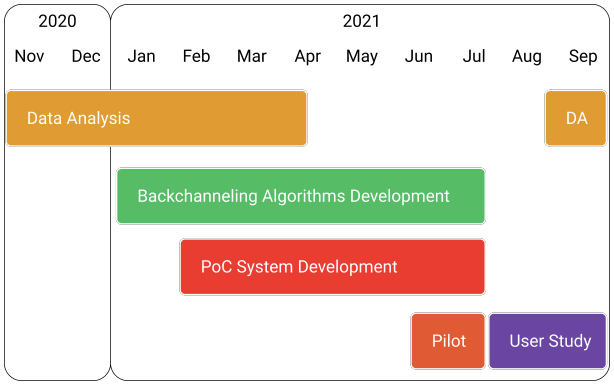


Figure 1: A timeline of research activities presented from November 2020 to September 2021.

perception of CAs by older adult users show positive results [40, 84, 105, 106]. Moreover, it has been proven that patients are more willing to share personal information with virtual therapists than human therapists [49]. Applying CAs to support caregiving to older adults also draws recent attention of the CHI community, resulting in a Special Interest Group (SIG) discussion in CHI 2020 [84].

However, adopting CAs to support health monitoring and disease screening is a relatively new trend [37]. For example, while tablet-based applications have been developed to assist clinicians to run tests for neurocognitive disorders (NCD), e.g. MoCA App [1], to the best of our knowledge, CAs for NCD tests have not yet been well-studied. Many open problems related to this domain of application exist [37], e.g., how to make the CAs can "behave" more like a human as it converses with the user [34, 84], as well as how age may impact the user experience in interacting with these CAs [37, 75, 93].

2.2 Handcrafting Backchanneling Rubrics

The term "backchannel" was coined and defined as the messages delivered by listeners in a conversation, without the intent to take a turn [102]. Bangerter and Clark [5] defined a backchannel as a listener's response that happens during the speaker's turn, without taking a separate turn. More recently, researchers categorized backchannels along different dimensions, namely, by content (non-lexical, phrasal or substantive) [33], by function (backward-looking or forward-looking) [96], by relation to the speaker's utterance (generic backchannels/assessments or specific backchannels/continuers) [6, 25, 79, 88] and by proactivity (reactive or proactive) [91].

From the perspective of proactivity, researchers studied the passive characteristics of backchanneling according to the unilateral view of conversations, which may also be referred to as "reactive tokens" [14, 103] or "response tokens" [24]. This perspective is known as the reactive backchanneling theory [91]. On the contrary, other researchers started to study the listener's active participation in the development of a conversation, which also extends to backchanneling [6, 10, 15, 61–63] – these studies form the proactive backchanneling theory [86]. We referenced these studies in stating the definition of reactive backchannels (RBC) and proactive backchannels (PBC) in the introductory section. Table 1 shows the connection between proactive perspective and other perspectives.

Research studying human backchanneling behaviors started with handcrafting rubrics for generating backchannels [95, 96]. Subsequent research referred to the speaker's utterances that trigger backchannels as "backchannel-inviting cues" [26]. Commonly known backchannel-inviting cues are acoustic features, such as pause and pitch (falling or rising slope) [26, 65, 92]. There may also be linguistic features such as a final Part-of-Speech bigram in "DT NN", "JJ NN" or "NN NN" [26]. Prosodic and linguistic features are usually combined to achieve a better result if manual transcription or transcriptions from automatic speech recognition (ASR) is available [26]. Visual cues such as gaze have also been taken into consideration [74]. Research interests in studying user experiences in rule-based backchanneling have expanded from a target audience being adults to specific age groups such as kids [69].

2.3 Model-based Backchanneling using Machine Learning Algorithms

Model-based backchanneling approaches divide the task into two parts: prediction and action, referring respectively to the prediction of opportunities for backchanneling from observing the speaker's behaviors and choosing the appropriate type of backchanneling [73]. Previous studies formulated the problem in different ways, e.g., focusing on prediction and choosing the same backchannels [34]; bundling multiple binary classifiers to predict different types of backchannels and giving corresponding actions [35]; training a multi-class classifier to predict and act at the same time [35]. Various machine learning algorithms have been applied, such as locally weighted linear regression [87], Hidden Markov Model (HMM) [58, 59], Support Vector Machines [52], Long Short-Term Memory networks [28, 34, 77, 78] and hybrid time-delay neural network (TDNN)/HMM system [65].

Feature engineering is another important component for ML-based methods. Similar to the acoustic backchannel cues in rubric methods, prosodic features are commonly used as model inputs. Prosodic features include Mel-Frequency Cepstral Coefficients (MFCC), pitch (fundamental frequency), energy, speaking pause (voicing probability) and pitch/power contour [34, 65]. Besides, fundamental frequency variation (FFV) , duration, Spectral Flux, and voice quality-related features could be taken into consideration as well

[1] https://www.MoCAtest.org/app/

Proactivity	Content	Context	Function	Description
Reactive	Non-lexical	Generic	Backward-looking	Vocalic sounds that have little or no referential meaning e.g. "mm hm", "uh huh", "yeah"
Proactive	Phrasal	Specific		Lexical expressions of acknowledgment and assessment e.g. "really?", "I see", "I know", "good", "fine", "okay"
	Substantive		Forward-looking	Follow-up questions or encouragements to ask the speaker to talk more e.g. "Anything else?", "Keep going"

Table 1: Comparison among various categorizations of backchanneling: Proactivity, Content, Context (relation to previous speaker's utterance) and Function [6, 25, 33, 79, 88, 91, 96].

[76, 78]. Currently, more comprehensive feature sets are available, such as the ComParE feature set [98], which is a growing set of acoustic features (6,373 in total) [98]. ComParE has been widely used in various acoustic recognition tasks, including emotion [80, 83], speaker [82], eating condition [81] and Alzheimer's disease detection [47]. This work is the among the first to adopt the ComParE features set for backchanneling prediction.

3 DATA ANALYSIS

To address the first research question – "**RQ1**: When do reactive and proactive backchannels happen in a task-driven conversation with older adults in Cantonese?", we tried to learn from real-world conversations between older adults and trained assessors – We analyzed a dataset of Montreal Cognitive Assessment (MoCA) recordings based on older adult participants having a conversation with human assessors. This MoCA dataset was collected by 6 experienced clinicians (5 females and 1 male) in 2 years, from June 2015 to July 2017. It included 246 Cantonese conversations between trained assessors and older adult participants (171 males and 75 females), each being approximately 30 minutes in duration, and with hand transcriptions aligned at the word-level with speech. The scale of this dataset was comparable with some of the largest datasets used for backchanneling studies, such as SwDA [65, 77, 78]. Participants in this MoCA dataset were aged between 77 and 94, with an average age of 82.9. Data analyses strictly followed the project's Institutional Review Board (IRB) approval to protect the privacy of participants. The tasks that participants performed in the MoCA dataset were listed in the Supplementary Material I.

3.1 Coding Process for Backchannels

To code the assessors' backchannels in this dataset, we followed these three steps:

(1) Inspecting the transcripts and identifying "backchannel words" related to reactive backchannels (RBC) and proactive backchannels (PBC). Similar methods of labeling backchannels in the corersponding textual transcripts had been used in previous work [78],
(2) Developing a coding scheme based on RBCs and PBCs, and improving inter-rater reliability (IRR) of that coding scheme, and
(3) Building rubrics to auto-coding backchannels on the whole MoCA dataset based on the coding scheme and the ground truth coded by human coders.

First, Researcher A with full professional proficiency in Cantonese inspected all the assessors' utterances of 110 conversations, with 15,455 unique utterances in total. The goal was to find "backchanneling words" which might be candidates of backchannels, such as "hmm" and "right". Researcher A followed the definition of backchannels as "not occuring in separate turns, but during the speaker's turn" [5] to identify utterances that encouraged the participant to talk more without interruption. The inspection resulted in 11 RBC words and 12 PBC phrases (see examples of RBCs and PBCs in Figure 2). Then Researcher A developed a coding scheme (see Table 2) – If the assessor's utterance was not considered as a backhannel, it would be coded as 0. Researcher A shared the coding scheme with Researcher B, and both started to code the backchannels on the word-level aligned transcripts of four conversations from the MoCA dataset.

Coding 4 conversations were divided into 2 rounds, with 2 conversations in each round. After the first round of coding, Researchers A and B reviewed all the codes and discussed inconsistent codes, finalized the ground truth and refined the coding scheme, and then conducted a second round of coding. Besides the content categorizations of backchannels, rubrics for coding the assessors' backchannels were also developed:

- (R1) A backchannel should follow an utterance of the participant (which included an utterance followed by a long pause), or a previous backchannel from the assessor.
- (R2) A backchannel should indicate that the assessor intended to hear from the participant instead of taking a turn turn.
- (R3) It did not matter whether a backchannel was followed by an utterance (successful backchanneling) or not (unsuccessful backchanneling).
- (R4) If some information was included in the backchannel, it should be intuitive and minimal given the interaction context, e.g. "minus", "aunt".

The results of two rounds of coding are reported in Table 3. The IRR (Cohen's Kappa) between Researchers A and B for the second round of coding was 0.803, which showed substantial agreement between coders [42]. After resolving the mismatches in these two rounds of coding, we obtained the ground truth of the backchannels in those conversations.

Given the large volume of assessors' utterances in the MoCA dataset, we tried to develop a rubric-based method to code the backchannels in 246 conversations automatically. Below are the rubrics we used to auto-code RBCs and PBCs.

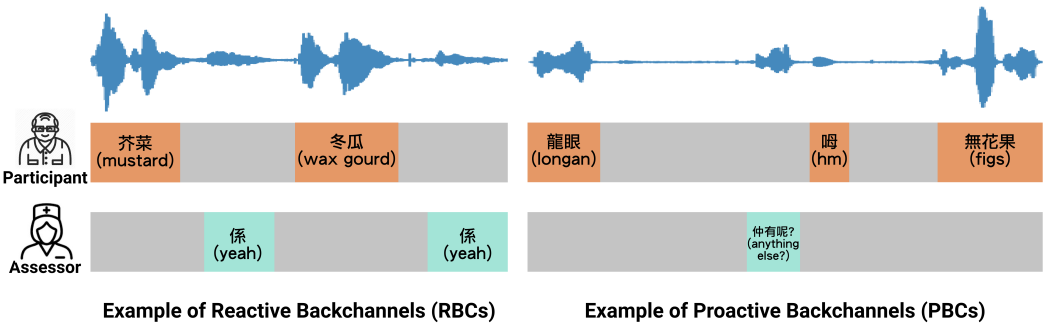


Figure 2: Examples of reactive backchannels (RBCs) and proactive backchannels (PBCs) from the MoCA dataset. The participant was conducting 1-min verbal fluency task, trying to say as many vegetable/fruit names as possible within one minute and receiving RBCs and PBCs from the assessor.

Category	Code	Definition	Examples
RBCs	1	Reactive backchanneling to show understanding and agreements	oh, hmm, ah
PBCs	2	Proactive backchanneling to encourage the speaker to speak more	keep going, anything else

Table 2: Coding scheme based on categorization of content in backchanneling.

	# of Utterances	# of Backchannels Coded	# of Consistent Codes (%)	IRR (Cohen's kappa)
Round 1	588	A: 46; B: 36	563 (95.7%)	0.676
Round 2	544	A: 29; B: 35	53 (97.8%)	0.803

Table 3: Results of two rounds of coding backchannels.

Rubrics to auto-code RBC:

- (RBC-R1) only one word in the RBC words, e.g. "hmm", "oh", "uh", "ah", "huh";
- (RBC-R2) at least 1000ms after the assessor's previous utterance and 1000ms before assessor's next utterance.

Rubrics to auto-code PBC:

- (PBC-R1) at most 8 words (the longest PBC phrase has 3 words), including PBC phrases e.g. "keep going", "next", "understand", "great", "awesome", "no rush", "anything else";
- (PBC-R2) at least 1000ms after the assessor's previous utterance and 1000ms before assessor's next utterance.

The lists of RBC words and PBC phrases with English translations provided in Supplementary Material II. The reason for adding RBC-R2 and PBC-R2 was to guarantee that the utterance was not a part of a longer utterance such as "Hmm...time is up", which might involve taking a turn. Our auto-coding scheme achieved substantial agreement with the ground truth (*Cohen's kappa = 0.774*) [42]. We used these rubrics to code all the assessors' utterances in the MoCA dataset, resulting in 2,732 RBCs and 2,037 PBCs.

3.2 Insights from Data Analysis

As stated in RQ1, we aimed to study the timing of backchanneling in task-oriented conversations. We planned to analyze coded backchannel occurrences to gain insights from real assessors' behaviors. To understand the timing of RBCs and PBCs after participants' speech utterances, we first drew a boxplot to show the distribution of time intervals between the end of participants' previous utterance and the beginning of backchannel (see Figure 3).

In Figure 3 we observed that assessor's RBCs tended to have shorter time interval and much smaller variance than those of the PBCs ($t = 26.18$, $p << 0.01$, after removing outliers over 2 SDs). This indicated that that RBCs were likely to occur after a certain duration of pausing that followed the participants' speech. In comparison, PBCs were likely to occur significantly later, and hence they did not serve as an immediate response to the speaker utterance. This observation aligned with our definition of RBCs – that they were generic and independent of other factors such as contextual information (e.g., the task, participant, etc.), and the timing of RBCs after the pauses were relatively stable ($M = 178.2, SD = 1335.7$). As for PBCs, their timing after the pauses were much longer with a larger variance ($M = 1285.1, SD = 2371.9$), which might reveal that there were factors other than pause duration that affected the timing of PBCs. Hence, we inspected the distribution of PBCs according to three other factors: within-task progress, between-task differences, and between-participant differences, and summarized our observations according to possible elements that might affect how proactive the human assessors may give PBCs.

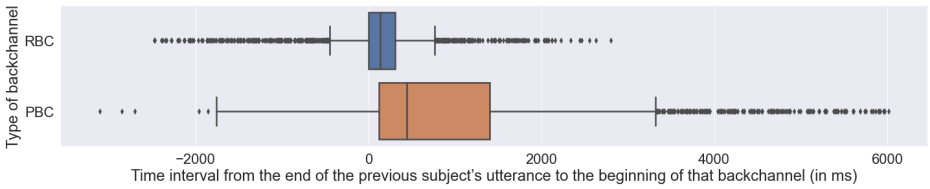
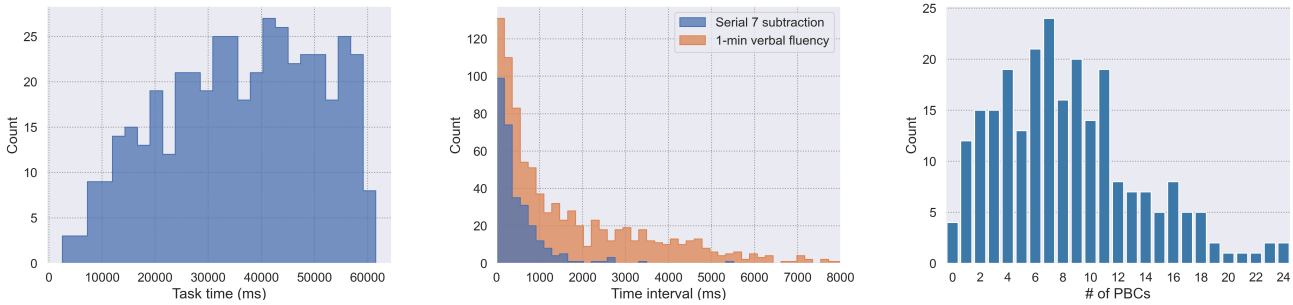


Figure 3: The assessor's PBCs may begin after a longer time interval from the end of the previous participant's utterance, as compared with the assessor's RBCs. The time interval may also be negative, which implies an overlap with previous participant's utterance.



(a) **(b)** **(c)**

Figure 4: The number PBCs given by expert assessors varied along 3 dimensions: within task (a), between tasks (b) and between subjects (c). Plot (a) describes the distribution of PBCs throughout the progress of the 1-min verbal fluency tasks, Plot (b) describes the distribution of time intervals and Plot (c) describes the distribution of the number of PBCs received by participants.

3.2.1 Within-task differences in PBCs given by expert assessors. We first investigated the timing of PBCs as a task progresses – Figure 4a plotted the distributions of PBCs among all the 1-min verbal fluency tasks. We observed that assessors gave more PBCs halfway into the task, and then the number fell before the task ended. One potential reason may be that the participant ran out of answers later in the task and assessor tried to use more PBCs for encouragement, while the assessor also noted the time limit and tended to give fewer PBCs when there were only a few seconds remaining. These results showed that assessors may vary their proactive levels dynamically according to task progression.

3.2.2 Between-task differences in PBCs given by expert assessors. We noticed the differences in the PBCs given across tasks in the MoCA tests. Considering the number of PBCs given, we identified three major types of tasks: Type I) *tasks requiring a one-off response*, where no PBCs were given, e.g., "please repeat the sentence…"; Type II) *tasks requiring a series of responses in a given time*, where several PBCs may be given, such as the 1-min verbal fluency task "please say as many animal names as possible in one minute", and the serial 7 subtraction task "please use 100 minus 7, and continue to subtract 7"; and Type III) *tasks requiring open-ended self-disclosure*, where few PBCs may be given, e.g., "where is your favorite place and why".

Even for the same type of tasks, such as the 1-min verbal fluency task and the serial 7 subtraction task, both of which required a series of responses from participants, there existed differences in the duration of tolerated silence before a PBC was given (see Figure 4b). Most PBCs for the serial 7 subtraction task occurred within 1 second after participants' previous answer, while for the 1-min verbal fluency task, the corresponding interval could be as long as 10 seconds. A possible explanation may be due to the level of difficulty of the task. For MoCA, many older adult participants stated that the serial 7 subtraction task was quite difficult for them. The expert assessors were mindful of the level of task difficulty and tended to give more PBCs to encourage the participant for more difficult tasks.

3.2.3 Between-participant differences in PBCs given by expert assessors. Figure 4c shows the total number of PBCs received by each participant in the MoCA dataset. Although each participant was asked to conduct the same set of tasks, the counts of PBCs received per person varied substantially from 0 to 24. This result suggested that there existed between-participants difference in how expert assessors gave PBCs.

4 METHODOLOGY OF BACKCHANNELING

Given the difference in functions and timing between reactive backchannels (RBCs) and proactive backchannels (PBCs), we developed two different models to generate these two kinds of backchannels. This section describes the models and discusses how they are integrated into a functional backchanneling system. To give an overview, when a speaking interval was detected, the previous speaking utterance followed by this interval would be segmented. Then, the speaking utterance was fed into OpenSmile feature extractor to obtain selected ComParE features, and those features

were used by the trained SVM model to make RBC predictions. For PBC prediction, three scores were independently calculated and a weighted sum of those scores was used to trigger PBCs. Participant Score was calculated by passing the speech of the exact same sentences to SVM model. In a certain task, Progress Score was updated as the task was going on, and when speaking interval occurs, Pause Score would be calculated as the length of pauses at that time stamp. The pipeline of processing speech data is illustrated in Figure 5.

4.1 Reactive Backchanneling Algorithm

Since RBCs usually follow the previous speaker's utterances as a direct signal of understanding or agreement [91], we pose the problem of finding RBCs as a binary classification problem – Given a speaker's utterance, the RBC prediction model will make decision regarding whether the utterance contains the backchanneling acoustic cues [65] to trigger an RBC. Building on this definition, we investigated and compared multiple features and models, then proposed a method leveraging the ComParE feature set, a LASSO-based feature selection algorithm, and SVM classifier to predict RBCs.

4.1.1 Feature Engineering. Instead of using existing features from other languages and dialects, we performed feature engineering to retrieve appropriate features from Cantonese utterances, since backchanneling features are largely affected by languages and cultures [14, 17, 29]. As introduced in 2.3, dominant acoustic features are extracted based on the ComParE feature set [81, 83], for triggering backchanneling, and all the spoken interactions were in Cantonese.

A LASSO-based algorithm was used to perform feature selection and reduce feature dimensions. The algorithm utilized L1 penalty on linear regression models to force the model to have a sparse weights distribution, by which the features with non-zero weights are selected. We noticed that the selection process was sensitive to the random initialization parameters. Hence, we further applied a stability selection algorithm [54, 85]: instead of using the result of one round of selection on the whole dataset, we used stability selection to randomly subsample half of the data for N rounds, aggregated all the features selected and recorded the times they get selected. Then a threshold was used to keep the most frequently selected features. In our experiments, this threshold was set to 0.6 as default. As a result, 34 out of 6373 features were selected as input features to train the RBC prediction model. In addition, we also used a prosodic feature set as a baseline, including the 13-dimensional Mel-Frequency Cepstral Coefficients (MFCC), fundamental frequency (F0), and sum square energy, which were commonly used in many previous works [34, 51, 65]. Selected ComParE features and prosodic features were listed in the Supplementary Material III.

4.1.2 Theoretical Experiments for Reactive Backchanneling. A series of experiments were conducted to obtain the best-performing configuration of features and models. We evaluated the selected ComParE features using the Multilayer Perceptron (MLP) and Support Vector Machine (SVM) models. A reference baseline experiment was conducted using prosodic features and the Long Short-Term Memory (LSTM) model, according to the state of the art [34, 65]. The reason that we did not combine the ComParE feature set with the LSTM model is that ComParE feature set did not include the time-domain information of speech segments, and hence the LSTM model cannot be used.

As for the training data, 2,732 RBCs coded in 3.1 were used. According to previous work [26], we considered the participants' utterances before coding the assessors' backchannels as *RBC cues*, which may be more likely to contain acoustic information for triggering RBCs The negative samples were randomly selected from the participants' utterances that were not followed by assessors' backchannels, referred to as *non RBC cues*. For a legitimate comparison, the sampling process took both participant and task distributions into consideration, i.e. if there were n RBC cues coming from the participant x in task t, then n non-RBC cues would be randomly sampled from the participant x in task t. In this way, we obtained a balanced data of RBC cues and non RBC cues.

From the experimental results (see Table 4), we observe that the results based on selected ComParE features and SVM generally outperformed the LSTM baseline, while the baseline method had a higher recall rate. As stated in the definitions of RBCs and PBCs, backchanneling is an optional behavior and it is not necessary to give a backchannel within a certain timing, so the recall metric may not be a major concern this context. Therefore, taking performance and inference speed into consideration, the SVM model with the ComParE feature set is selected to be the best one for implementation in our system.

4.2 Proactive Backchanneling Algorithm

The function of PBCs is to encourage speakers to continue talking, and the most intuitive way of triggering PBCs is to take place after the users have stopped speaking for a while. More specifically, PBCs tend to be triggered by long pauses instead of speech segments and short pauses. This implies that the methodology for predicting RBCs may not be applicable to predicting PBCs. Moreover, as discussed in 3.2, trained assessors were observed to take task progress, type of tasks, and characteristics of participants into consideration when giving PBCs. Hence, we introduce a comprehensive scoring method, *Triple P Scoring Method*, to imitate those adaptive PBC strategies. The *Triple P Scoring Method* included three main components: ***P**ause Score*, ***P**rogress Score* and ***P**articipant Score*. We calculated the final PBC score through a weighted sum of those three scores, and a PBC will be triggered if the PBC score exceeds a threshold. The threshold was determined by the data collected through piloting our system with 10 participants. Below we introduce how we calculated those three scores for Type II tasks (e.g. 1-min verbal fluency task and serial 7 subtraction task), where most PBCs occurred in the MoCA dataset.

4.2.1 Pause Score. As mentioned earlier, pausing is the most intuitive cue for providing PBC, so the Pause Score is a critical component for triggering PBC.

First, a Log-normal distribution was selected through distribution selection with minimum sum square error in the MoCA dataset, as a Probability Density Function (PDF) to model the distribution of the intervals between PBCs and their related speaker utterances (see Figure 6). Then, we calculated the Cumulative Density Function (CDF) of the modeled PDF, and the value of CDF at a certain point of time was used as the Pause Score of that moment. According to

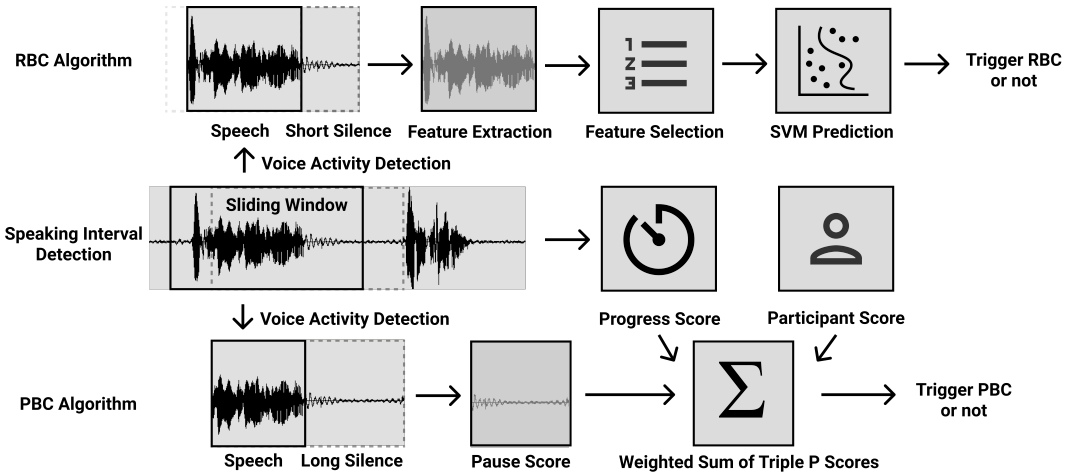


Figure 5: Overview of the pipeline for processing speech data and generating RBC and PBC decisions.

Feature	Model	Accuracy	Precise	Recall	F1
Prosodic	LSTM	0.615/0.638	0.573/0.610	**0.866/0.859**	0.689/**0.714**
ComParE	MLP	0.650/0.642	**0.670/0.671**	0.579/0.623	0.615/0.646
	SVM	**0.692/0.655**	0.656/0.646	0.793/0.760	**0.718**/0.695

Table 4: Experiment results on RBC prediction models. Metrics are denoted as Cross Validation/Test.

the CDF, the obtained Pause Score was in the range of (0, 1) and increases as the participant's pause becomes longer.

The distribution of the interval between participant's utterances and PBCs is:

$$PDF_{lognorm}(t_{pau}; z, s) = \frac{1}{sz\sqrt{2\pi}} exp(-\frac{log^2(z)}{2s^2}) \tag{1}$$

$$z = \frac{t_{pau} - \mu}{\sigma} \tag{2}$$

where t_{pau} means the silence time of participant, and μ, σ and s are parameters of log-norm distribution estimated by maximum likelihood from the MoCA dataset. The Pause Score could be expressed as:

$$Score_{pau} = CDF_{lognorm}(t_{pau}; z, s) = \int PDF_{lognorm}(t_{pau}; z, s)dx \tag{3}$$

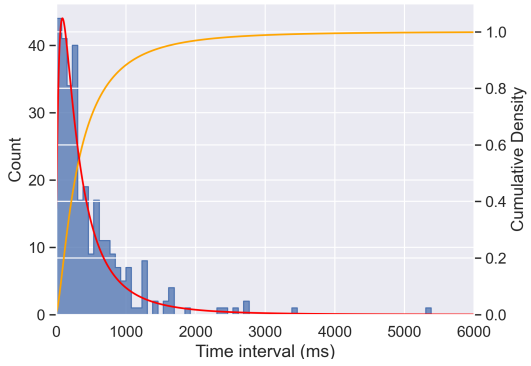
Besides, as shown in 3.2.2, the s of the intervals between the speaker's utterances and the PBCs are largely depending on tasks. For example, it seems that assessors tended to have lower tolerance of silence and gave more encouragement to the older adult participants whey they were undertaking relatively harder tasks, e.g. serial 7 subtraction. Hence, we differentiated CDFs and PDFs for different tasks. For instance, Figure 6 compares two tasks: the serial 7 subtraction task and 1-min verbal fluency task. We can see that the Pause Score of the serial 7 subtraction increases more rapidly than that of the 1-min verbal fluency task, which means that algorithm tended to provide more PBCs in serial 7 subtraction than in 1-min verbal fluency task.

4.2.2 Progress Score. As discussed in 3.2.1, there was a clear difference in the number of PBCs provided by the assessors as the task progresses. For example, the assessor would provide more PBCs during the second half of the one-minute naming task. We model this behavior of the assessors with the Progress Score. Similar to Pause Score, a PDF obtained from distribution selection was used to model the distribution of PBCs during the task period. Moreover, to derive a computable score, we discretized the PDF into Probability Mass Function (PMF) in bins of 100ms, and then used a scaling factor to rescale the maximum of PMF to 1, by which we obtained a score in [0, 1]. Based on the MoCA dataset, the Skew-Normal distribution was selected as PDF, and this process can be described by the following function:

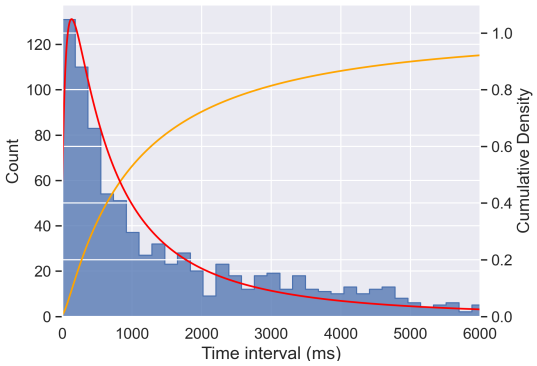
$$Score_{pg} = k * PMF_{skewnorm}(t_{task}) \tag{4}$$

Where $Score_{pg}$ stands for Progress Score, k is the scaling factor, and t_{task} is the time that a task has been proceeded. The PDF fitting result is shown in Figure 7. Similar to our observation in the MoCA dataset, Progress Score peaked in the second half of task progress.

4.2.3 Participant Score. Participant Score was a score generated at the beginning of a test, describing the proactivity level of backchanneling for each participant i.e. how difficult to trigger a PBC. Participant Score was motivated by the finding in 3.2.3 that there was a clear individual difference among participants. To simplify this question, we considered it as a probabilistic classification task, where the input was a segment of speech from the participant with fixed content and the output was a score. We first used SVM to classify participants into two classes: participants received more PBCs and participants received less PBCs. Then the output d of SVM,



(a)



(b)

Figure 6: Distributions of pauses before PBCs show a difference between two tasks: a) serial 7 subtraction task b) 1-min verbal fluency task, where assessors might have a lower tolerance of silence and were more inclined to give PBCs faster in a) compared with b). Probability Density Function is in red and Cumulative Density Function is in orange.

i.e. distance from input data to SVM classification hyperplane, was used as a classification score. Next, Platt Scaling [71] was applied on d to obtain a probabilistic value with range of (0, 1) from the output of SVM, denoted as Participant Score, through the following function:

$$Score_{subj} = Platt(y = 1|d) = \frac{1}{1 + exp(\alpha f(d) + \beta)} \quad (5)$$

where $Platt(y = 1|d_{subj})$ refers to Platt Scoring, d refers the output of classification model, α and β are parameters learnt from SVM training data [71]. Table 5 shows the results of theoretical experiment using SVM to classify two groups of participants.

4.2.4 Overall PBC Score. To summarize, Pause Score measured utterance-level timing, telling when PBCs should occur after participant's utterance, while Progress Score measured task-level timing, indicating when PBCs should occur within a task. Participant Score adjusted the proactive level to adapt to different participants. Then

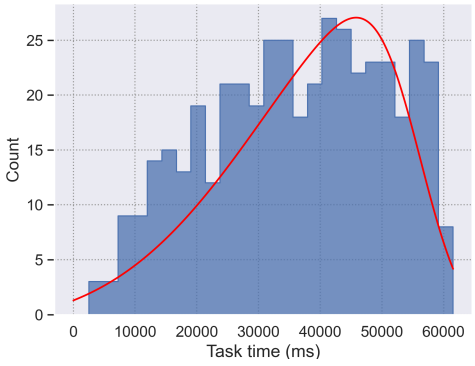


Figure 7: Fitting result of Probability Density Function (red) for PBCs generated through the progress of 1-min verbal fluency task. It shows the frequency of PBCs given increases as the task proceeds, and decreases when the task is coming to an end.

SVM	Acc	Pre	Recall	F1
Valid	0.623	0.611	0.624	0.605
Test	0.571	0.600	0.576	0.589

Table 5: Results of SVM classification on two classes of participant.

we calculated an overall PBC score through combining those three sub-scores with a weighted sum:

$$Score_{PBC} = w_{pau} * Score_{pau} + w_{pg} * Score_{pg} + w_{pt} * Score_{pt} \quad (6)$$

For the convenience of tuning parameter, we limited the sum of w_{pau}, w_{pg} and w_{pt} to be 1, of which the range of $Score_{PBC}$ was (0, 1). Then we set a threshold thr_{PBC} to make PBC decisions if the $Score_{PBC}$ was beyond that threshold. The lower thr_{PBC} was, the more likely it was for users to receive PBCs. In order to find optimized setups for user experience, the hyper-parameters thr_{PBC}, w_{pau}, w_{pg} and w_{pt} were tuned according to statistics and user feedback from piloting our algorithm with ten participants. For instance, w_1 could be turned down to increase the response time of PBC; thr_{PBC} could be turned up to reduce the number of PBCs received by users.

4.3 Implementation of Backchanneling Algorithms

We integrated the two models into a fully functional system, which analyzed a user's speech data in real-time and provide RBCs and PBCs whenever appropriate.

Previous studies mainly considered a continuous prediction, i.e. continuously feeding input audio into a prediction model and deciding whether to give a backchannel at each frame [34, 65, 78]. Voicing Probability was used as input feature in those methods, with an expectation that the prediction model was able to learn the correlations of backchannels and speaking intervals/utterances. However,

the uncertainties introduced by prediction models brought a risk of interrupting speakers in case of false alarms .

In our system, instead of using Voicing Probability as a feature, we independently designed a Speaking Interval Detection (SID) module to recognize speaking utterance and speaking interval from the input audio. First, Voice Activity Detection (VAD) was adopted to pre-process the input audio with a sliding window, aiming to label input frame as voiced or unvoiced frame. Next, in order to aggregate voiced/unvoiced frames and avoid noising prediction spikes, a delay trigger mechanism was used to further process VAD results, i.e. the audio segments were considered as *speaking utterance* only when the number of speech frames detected was beyond a threshold; similarly, the segments would be considered as *speaking interval* when the number of silence frames was beyond a threshold.

With this method, speaking boundaries were visible to our system and RBC/PBC modules would only be triggered by speaking intervals. Besides reducing the risk of interrupting the speaker aggressively, this pre-processing method could lower the computation cost as well. It only passed speech data to backchannel prediction models when speaking intervals were detected rather than triggers those models recurrently. It could reduce the computation latency of our backchanneling pipeline, which was an important consideration of making backchannel decisions in real time.

5 EVALUATION OF PROOF-OF-CONCEPT SYSTEM

To answer "**RQ2**: How do reactive and proactive backchannels provided by a task-oriented CA affect its conversation with older adults in Cantonese?", we developed a proof-of-concept system *TalkTive* for speech-based NCD screening, with a multimodal interface and with the RBC and PBC modules embedded. Then we conducted a between-subject study to evaluate the performance and user experience of our *TalkTive* system. The full system (Condition 2) was used to compare two conditions: Condition 0 (baseline condition) with the preset task functions and no backchanneling, and Condition 1 with the same task functions but providing RBCs. We obtained institutional IRB approval for the whole project prior to the study.

5.1 System Architecture and Interaction Flow

The *TalkTive* system consisted of two parts: a React frontend as graphical user interface (GUI, see Figure 8) and a Flask python server (see Figure 9). Users interacted with the GUI to complete a series of MoCA tests as instructed by the conversational agent. In this process, the GUI issued corresponding commands to the backend as API calls, such as to open or close the RBC and PBC modules, and sent information inputted into the interface e.g., user's age to the backend. It also played back the speech output generated by the server.

More specifically, when a user came to the page of a new task, *TalkTive* would introduce the task by playing an audio clip of task description. The user could replay the recording or start the task. For tasks with a time limit, a countdown would appear after the user clicks the "Start answering" button. While the user was providing speech responses to the given question, *TalkTive* ran the algorithms in Section 4 to provide RBCs and PBCs in real time.

To add the action module to the *TalkTive* system, we invited a native Cantonese speaker experienced in conducting MoCA tests to record a Cantonese backchannel library. To build this library, we selected the most common RBCs and PBCs coded in the MoCA dataset, segmented related audio clips (which may also contain participant's speech and background noise), and passed them to an experienced MoCA assessor as prompts for audio recording. The assessor mimicked those audio clips of backchannels in a sound proof room. We indexed this Cantonese backchannel library and integrated into our backchanneling pipeline. If a RBC or a PBC was triggered, *TalkTive* system would randomly play a piece of corresponding backchannel audio from the Cantonese backchannel library as feedback.

5.2 Tasks

There were three kinds of MoCA screening tasks, categorized based on the type of expected answer 3.2.2 – Type I questions asking for a one-off answer (e.g. "please repeat the sentence…"), Type II questions prompting for a series of responses (e.g. "please say animal names as many as possible in one minute"), and Type III open-ended questions. To ensure representativeness of experience in the user study, we sampled the three types of tasks (at least one question from each type) according to their frequency in MoCA. This resulted in a trial MoCA test of nine tasks: three Type I questions, five Type II questions, and one Type III question as listed below:

- (T0) *Type I - Sentence repetition*: Please repeat the sentence that I said: [tongue twister in Cantonese]. (Trial task)
- (T1) *Type II - 1-min verbal fluency*: Please say fruit names as many as possible in one minute.
- (T2) *Type I - Sentence repetition*: Please repeat the sentence that I said: [tongue twister in Cantonese].
- (T3) *Type II - 1-min verbal fluency*: Please say animal names as many as possible in one minute.
- (T4) *Type II - 1-min verbal fluency*: Please say vegetable names as many as possible in one minute.
- (T5) *Type II - Serial 7 subtraction*: Please begin with 100 and count backwards by 7.
- (T6) *Type II - 1-min verbal fluency*: Please say place names in Hong Kong as many as possible in one minute.
- (T7) *Type III - Open-ended self-disclosure*: Could you please share a place you like and why?
- (T8) *Type I - Understanding*: How to say the word (clap) and how to perform that action?

In particular, T0 was designed to collect participants' speech and generate Participant Score, which was be used in the subsequent tasks. The method used to generate Participant Score was described in 4.2.3. T0 also served as a practice task to familiarize the participant with the *TalkTive* system.

5.3 Participants

To conduct a between-subject user study with our target users – older adults, we recruited $n = 36$ participants (19 females and 17 males, aged from 61 to 84 with an average of 72.4), 12 for each condition. We asked all participants to complete a pre-study survey based on their experiences in using electronic devices. According

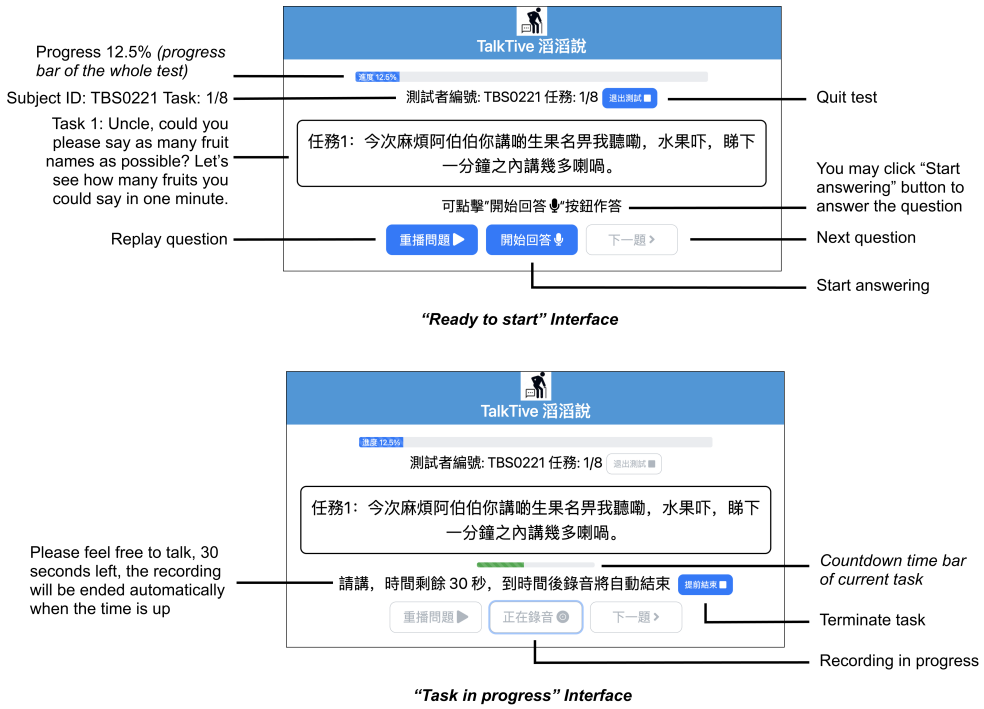


Figure 8: *TalkTive* interfaces: *"Ready to start" Interface* (above), where the participant received task instructions and got prepared for answering; and *"Task in progress" Interface* (below), where the participant provided answers in speech and could receive system-generated backchannels (Condition 1 & 2).

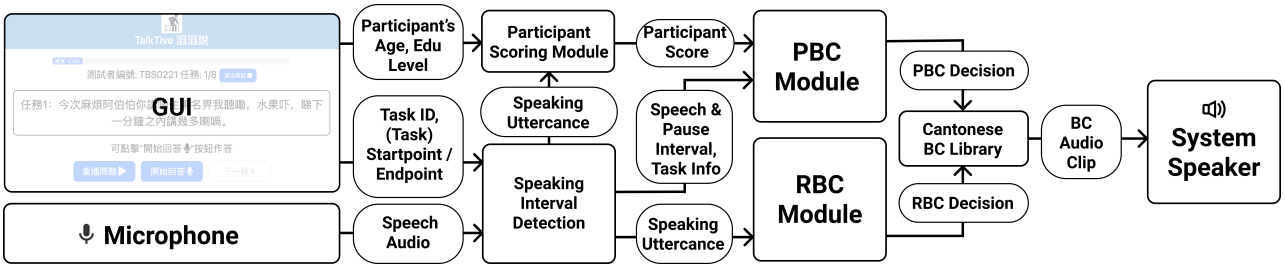


Figure 9: The backend of *TalkTive* system to predict RBCs and PBCs during conducting cognitive assessments.

to their replies, all participants had a smartphone, 13 (36.1%) had a tablet, and 13 (36.1%) had a laptop or a desktop or both (one may own multiple devices). Most participants reported using electronic devices 3 – 10 hours (17) and 1 – 3 hours (12) per day. Only three participants (8.3%) reported using electronic devices for less than one hour per day. These data suggested that our participants had daily access to electronic devices that could be used for CA-based NCD screening.

5.4 Study Procedure

The language used throughout the whole user study including the trial MoCA test was Cantonese. A researcher who was a native Cantonese speaker served as the experimenter and moderated the entire study process. At the beginning of the user study, the experimenter informed the participants that they would conduct a test of cognitive ability and memory. All participants signed a consent form, agreeing to join the research study and be audio recorded. After finishing a pre-study survey on participants' experience using electronic devices, participants were asked to watch a tutorial video of how to use the GUI as illustrated in Figure 8 by showing the steps to finish an example task (not used in the main study). After that participants were required to finish eight tasks independently, and the experimenter would not intervene unless there was a system breakdown. Upon completion of all eight tasks, the participants proceeded to fill out a post-study questionnaire on the same screen with assessor's facilitation (details in 5.5.2). They were required to rate their level of agreement to eight statements regarding their experience of talking to the (*TalkTive* system or baseline system without backchannel), according to a 7-point Likert scale. We concluded the study with an exit interview of four questions about

their user experience. We present the detailed survey and interview questions in the next subsection.

5.5 Measures

To evaluate the effect of backchannels provided by *TalkTive* and user experiences of using the system for speech-based NCD screening, we collected a series of measurements from both experienced human assessor's and older adult participant's perspectives.

5.5.1 Assessor's Validation of System-Generated Backchannel Responses. We audio-recorded all the conversations between the participants and the *TalkTive* system and marked all occurrences of backchannels generated based on system logs. We invited a trained assessor who has earned a certificate of running the MoCA test and also a native Cantonese speaker to evaluate the appropriateness of all backchannel instances given in Conditions 1 and 2. The assessor listened to all task recordings and was prompted to rate each instance of backchanneling with binary choice – inappropriate versus appropriate. This evaluation may be regarded as testing the precision of our backchanneling algorithms. Because of the optional nature of backchanneling [32, 96], the selection and placement of backchannels may vary according to individualized preferences. Hence it was impractical to obtain the ground truth of all possible backchannels and identify the false negatives, so we did not include recall in the analysis.

5.5.2 Post-study Questionnaire of Participants. To obtain the participants' subjective feedback based their experience in interacting with our system, we adopted eight statements from [17]. The participant needed to provide a rating those statements (five from a positive perspective and three from a negative perspective) using a Likert scale from 1 (strongly disagree) to 7 (strongly agree). The five positive statements were "the person who just asked me questions showed me that she understood what I said", "... she listened attentively to what I said", "... she encouraged me to talk", "... she was polite" and "the test went smoothly". And the three statements from a negative perspective were "... she seemed impatient", "... she seemed cold and unfriendly" and "... she interrupted me". All participants were asked to provide a rating those eight statements.

We adopted the eight statements to gain a comprehensive understanding of the participants' experience with the *TalkTive* system. We randomized the order of presenting the positive and negative statements to prevent the participants from simply giving the same rating to all items. To facilitate older adult participants to finish the post-study questionnaire thoughtfully, the experiment would read out each statement if participants had difficulty viewing the text by themselves. We also ensured that the participants fully understood the meaning of the scale. In addition, we invited the participant to give the reasons behind their ratings through thinking out loud, and we audio-recorded their verbal explanations.

5.5.3 Semi-structured User Exit Interview. After each participant finished the test and the post-study questionnaire, the experimenter conducted a semi-structured exit interview and audio-recorded the session with the participant's consent. The user interview has three main questions as listed below.

- Q1: How was your experience using this system? *Follow up:* How did you feel about communicating with the person (the voice) who just asked you questions?
- Q2 [Asked in Condition 1 & 2]: Did you notice the response given by the system, like "hmm", "yeah"? How did you feel about them? Why? *Follow up:* Do you think those responses were different from those given by humans? If so, what was the difference?
- Q3: Could you please give some suggestions to improve this system?

6 RESULTS FROM THE USER STUDY

In this section we report on the quantitative and qualitative findings of our user study, covering the performance of the proposed backchannel-generation algorithms for NCD assessment, and the user perception of and experience with backchannel-enabled CA assessors to those without this mechanism.

6.1 About 89% of System-Generated Backchannels were Validated as Appropriate by Trained Assessor

Examples of the timing of RBCs and PBCs in conversations were shown in Figure 10. To verify that the timing and form of the BCs generated during the study conform to the common practice of human assessors, we invited an expert to classify each BC instance based on perceived appropriateness. After we collected all the data of participants using the *TalkTive* system, a trained assessor coded all 649 BCs (518 RBCs, 131 PBCs) given to the 24 participants in Condition 1 and 2 (details see 5.5.1).

Participants in Condition 1 received a total of 267 RBCs (22.3 RBCs per test). Although more RBCs were generated by our system compared with the MoCA dataset (11.1 RBCs per test), most RBCs generated were considered as appropriate by expert: only 26 (9.7%) of them were coded as inappropriate. A total of 251 RBCs and 131 PBCs were produced in Condition 2 (20.9 RBCs and 10.9 PBCs per test). The number of RBCs given was slightly lower than that of Condition 1, and the number of PBCs was comparable with the number of PBCs given in the MoCA test (8.28 PBCs per test). Among all BCs in this condition, 46 instances (12.0%) were coded as inappropriate (see Table 6 for details), which demonstrated that our evaluator deemed the mechanism to be generally acceptable. It seemed that the provided PBCs had a slightly higher chance than RBCs to be regarded as inappropriate, which aligned with the proactive nature of PBCs that they were encouraging but had the risk of being aggressive.

To further investigate the causes of each inappropriate backchannel, we asked the assessor to take note of the reason behind her judgement. Overall, she specified three kinds of major causes (see Table 6). 1) "Not reply with a meaningful speech utterance". This mainly occurred when RBCs responded to the speakers' non-lexical utterances such as "uhmm..." which did not include meaningful answers to acknowledge or agree with. 2) "Urged the speaker", which happened mostly when multiple PBCs were triggered within a short period of time. Also, 3) "Interrupted speaker's thinking", which frequently occurred for both RBCs and PBCs in the serial 7 subtraction task.

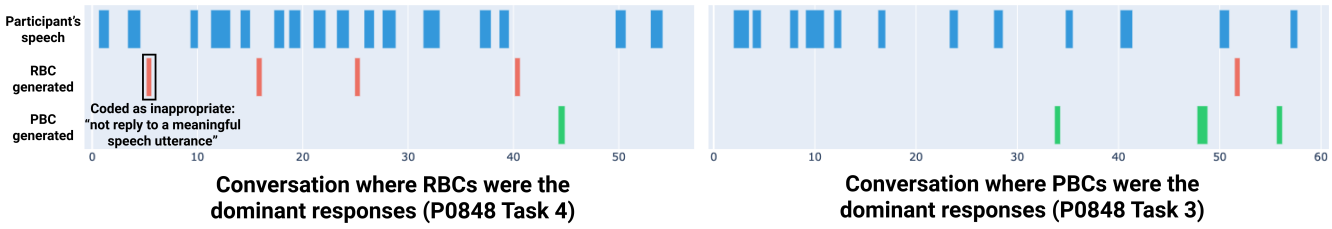


Figure 10: Examples of conversation where RBCs were the dominant responses (P0848 Task 4) and conversation where PBCs were the dominant responses (P0848 Task 3). We could see that most RBCs directly follow speakers' utterances while PBCs mainly occur in longer pause. The first RBC in P0848 Task 4 was coded as inappropriate because it did not reply to a meaningful speech utterance. All the other system-generated backchannels were coded as appropriate.

Con	Type	# of BCs	# of Inappropriate BCs (%)	Causes of Inappropriate BCs
1	RBC	267	26 (9.7%)	"not reply to a meaningful speech utterance" (13/26); "interrupted speaker's thinking" (10/26) "interrupted speaker's talking" (1/26) "quicker than the participant's response" (1/26) "overlapped with 'time's up'" (1/26)
2	RBC	251	27 (10.8%)	"not reply to a meaningful speech utterance" (24/27); "interrupted speaker's thinking" (3/27)
	PBC	131	19 (14.5%)	"urged the speaker" (10/19); "interrupted speaker's thinking" (8/19); "not reply to a meaningful speech utterance" (1/19)

Table 6: Results of assessor's validation of backchannels given in Condition 1 & 2.

When reviewing the occurrences of inappropriate backchannels in the audio recordings, we found many cases happened while older adults were mumbling to themselves. For "not reply to a meaningful speech utterance", the muttering of older adults might be detected by the Speaking Interval Detection (SID) module and considered as a speaker utterance and then trigger RBCs. Figure 10 shows a example. A similar situation may be observed for incomplete utterances with a long pause inside, which appeared frequently in the serial 7 subtraction task, as it was found to be difficult for the older adult participants. For example, the participant might get the first digit much earlier than the second digit, such as "eighty- [a long pause] six". Human assessor would know that "eighty" was a unfinished answer and would not backchannel to acknowledge that, while the pause might trigger our algorithm based on acoustic features to give a RBC. These kinds of situations indicated the need to use Automatic Speech Recognition (ASR) and Natural language processing (NLP) to understand the meaning of speakers' utterances, to identify the mumbling and also incomplete expressions, in order to yield more reasonable backchannels. However, currently there is a lack of high-performance ASR systems for older adult speech in Cantonese. In other cases, older adults may murmur as they were thinking. The SID may fail to capture such speech, and they system may treat it as a long silence and evoke PBCs, which may be considered considered intrusive for the speaker.

6.2 System with Proper Backchanneling Feature were not Perceived as Disturbing by Older Adults

To assess participants' subjective feedback on using our system, we compared the user ratings of eight statements in the post-study questionnaire across the three conditions. Because the answers of the post-study questionnaire were in the form of Likert-scale input, the data should be treated as ordinal measurements [11]. Hence we used the median to describe their central tendencies and applied Kruskal-Wallis non-parametric Test to evaluate the difference.

The statistical results were reported in Table 7. In general, there was no significant difference regarding agreements on the eight statements among three conditions. Although the results did not show that the system with backchannels outperformed the baseline condition in terms of encouraging speaker or listening to speakers attentively, they suggested that our system did not induce negative perceptions such as being less polite/patient/unfriendly after incorporating additional reactive or proactive backchanneling responses. However, regarding the central tendencies, Condition 1 with only RBCs has the lowest ratings for five statements, indicating that RBCs only may tend to perform worse than no BCs or both RBCs and PBCs.

Statements	Con 0	Con 1	Con 2	*p-value*	*H-value*
Positive Statements					
She understood what I said.	6.0	5.0	5.0	0.23	2.90
She listened attentively to what I said.	6.5	5.5	7.0	0.33	2.20
She encouraged me to talk.	6.5	6.0	7.0	0.53	1.26
She was polite.	7.0	7.0	7.0	0.42	1.74
The test went smoothly.	7.0	5.5	7.0	0.27	2.62
Negative Statements					
She seemed impatient.	1.0	2.5	2.5	0.07	5.43
She seemed cold and unfriendly.	1.0	1.0	1.0	0.11	4.45
She interrupted me.	1.0	1.0	1.0	0.10	4.55

Table 7: Medians and chi-square test results of users' level of agreement to these statements on a Likert scale of 7 (Strongly Agree) to 1 (Strongly Disagree) for Condition 0, 1 and 2. No significant difference was observed for all the statements among three conditions.

6.3 Qualitative Feedback on How Older Adults Perceived Backchannels

In this section of qualitative analysis, the goal was to understand the participants' thoughts and comments on the backchannels. Condition 0 has no backchanneling, and hence we report on the users' feedback for Conditions 1 and 2. there were no backchannels generated in Condition 0, here we mainly report the users' feedback on Condition 1 and 2.

6.3.1 Older adults reported that receiving only RBCs was not as good as gaining responses from human. More than half of participants in Condition 1 (7 out of 12) reported that the responses generated by the system were not as good as those generated by humans. Four participants said that they did not even notice the responses generated by the system (P0877, P0316, P0304, P0215). P0316 stated that the RBCs were not articulate enough:

> "I didn't notice that it was giving response to me. (actually 31 RBCs were given to her) Just asking me to finish one question and then next... At the beginning when I heard 'hmm', I had no idea about what it was doing. Really no idea. While after I heard two of that, I got to know that it was replying to me...Real human will be much warmer. I thought it (the system) was just a computer, a computer which could ask questions. It's different from face-to-face communication of human." (P0316)

It was reasonable for older adults to be insensitive to RBCs due to feeling nervous about the test, focusing on giving answers, even experiencing cognitive or hearing impairments. In other cases, although having noticed the existence of RBCs, P0215 stated that RBCs were rigid instead of human-like, and expressed a preference of in person communication:

> "There was no feedback. I noticed responses like 'hmm', but I knew it's just recording. I didn't think it's real. Human assessors would be better. For human assessors I could see their facial expressions to know whether they were paying attention." (P0215)

Other participants explained why they thought human assessors would outperform that system regarding responses given: "the responses generated by the system would be better than an inexperienced volunteer, but worse than an experienced assessor" (P0248); "human assessors would give more explanations...had more interactions" (P0235); "the responses given by humans would be clearer than those generated by the system" (P0280); "the system was less friendly compared with human" (P0874) or simply "the system had a lack of something compared with human" (P0877).

For the remaining five participants in Condition 1, only P0306 said the responses generated by the system were better than those from humans because some people may not give any responses and just wait: "The response was pretty good. It indicated that my answers were acceptable. The responses generated by the computer were better than those of human. Human would not give as much as response - he or she might just wait for you. Human might even not give response to you. Just let you think alone." The other four participants said that the responses generated were similar to human's without giving further explanation.

6.3.2 PBCs were appreciated by older adults, especially when they ran out of answers and were about to give up . Users gave more positive feedback on Condition 2 and they appreciated the existence of PBCs. In Condition 2, 8 out of 12 participants expressed that PBCs were well received by them. For example, P0250 found that PBCs were quite encouraging when she ran out of answers:

> "(The sound of) AI was quite machine-like, while this (system) was not. (I) think it was operated by a human. For example, just now when I had not answered, it was listening to me carefully, and encouraging me in the meantime, just like doing Q&A with you in person. It felt like that. We were nervous (in this kind of assessment), and it seemed that my mind suddenly went blank just now. (My brain) suddenly stopped (thinking). Could not think anymore. But it said 'feel free to keep thinking'. It listened to me quite attentively." (P0250)

In Condition 2, it seems that all participants noticed the system-generated backchanneling, in contrast with Condition 1 where four participants stated that the system-generated responses went unnoticed. One participant, P0221, even compared system-generated response with the audio recordings of task statements, and she

found the responses sounded more natural than task statements recorded by the human assessor. She stated the audio recordings of statements were machine-like, but the responses were not. She said, "While I was thinking, (it said) 'uhm', 'keep going', so I would really think (about the answers), much better than if it didn't give any response...because it made me feel like that there was really a person there, 'take your time', not like facing a cold computer".

When the participants were asked about the difference between the responses provided by the system compared with those from a human, P0232 said that the response was a good signal of listener's attentiveness, which might not even be provided by human. For example, P0232 and P0854 claimed that the system performed better than human and they felt more comfortable in talking to the system than with a human.:

> "(The responses from the system) showed that it valued my answers. Sometimes a person said 'you speak', but there was still no response even after I finished. It seemed that he didn't like that you spoke to him, didn't want to listen to you. If it had response, (it meant that it) paid attention to your talk." (P0232)

> "The responses were good. It would not make you feel nervous... Responses from a real human made me felt more nervous and stressed... (I was) more comfortable to talk to a computer (than a human)... The response from the system was encouraging you to keep talking, while the response from a system was more inflexible compared with human." (P0854)

7 DISCUSSION

7.1 Result Summary

In general, there was no significant difference among conditions with system-generated backchannels and the baseline condition with no backchannel given in either a positive or a negative way. This may imply that the participants did not feel overly anxious after receiving system-generated backchannels, even though some instances might be deemed intrusive or aggressive from a professional assessor's point of view. Clearly providing appropriate backchannels is still far away from providing "optimal" backchannels to help participants stay in the optimal arousal level, while eliminating the portion of inappropriate backchannels may be a concrete next step.

We found that reactive backchannels (RBCs), of which 89.8% were coded as appropriate by expert, may be ignored by older adults or considered as rigid. On the other hand, although proactive backchannels (PBC) have a higher risk of being perceived as pushy or intrusive by the expert (14.5%) than RBCs (10.2%), they were well received by most older adult participants in the given task setting. We observed in our study that older adults, especially those who were experiencing a decline in cognitive ability, expected the system to provide articulate and noticeable instructions and responses to guide them through the tasks. This preference of older adults was reflected as a general acceptance of PBCs. As shown in the user interview, older adults tended to consider receiving PBCs as being encouraged instead of being urged. Particularly, PBCs could be more helpful when participants intended to give up early, which was common in a cognitive assessment setting.

7.2 Design Considerations

Based on the above findings, we propose a set of design considerations for future improvement of conversational agents (CAs) conducting cognitive assessments with backchanneling function. First, we need to ensure that the generated speech of the task-oriented CA fits the content and nature of the tasks specifics, i.e., administrating a neurocognitive disorder screening test. As postulated by the well-known Yerkes–Dodson law [101], task performers should have an optimal mental arousal to perform a certain task. If the arousal of the task performers is lower or higher than the optimal level, their performance may be hindered by inactivity or anxiety. To guarantee that the participants' performance may fully reflect their actual cognitive abilities in the MoCA test, we hope to situate them in an appropriate arousal level and engage them in the task without making them feel too stressful. As shown in the qualitative results of the user study, PBCs were deemed effective by over half of participants of Condition 2 in keeping them engaged. In comparison, only a single participant responded as such for RBCs. Participants also expressed that PBCs were showing attentiveness instead of stressing them out. In the qualitative feedback. Hence, it may be beneficial to include PBCs to boost the task performers' arousal level so that they stay active in the process. At the same time, as suggested by the clinical assessors, we need to control the frequency and intensity of PBCs to avoid pushing the users beyond their normal arousal stage, which may been undesirable for cognitive assessment.

Second, for improving the precision of system-generated BCs, reducing the number of inappropriate PBCs may be prioritized because PBCs are more noticeable and might affect participants' arousal more significantly than RBCs. According to the Yerkes–Dodson law, difficult or intellectually demanding tasks may require a lower level of arousal to facilitate concentration for optimal performance [101]. This may explain why PBCs in the serial 7 subtraction task are often coded as inappropriate by the expert (10/19). To resolve the tension between encouraging versus pushing the participants while they perform intellectually demanding tasks, we may further investigate task-related adaptivity. Based on analysis of the MoCA dataset and results of user study, we may extend the cooling time of providing continuous PBCs in those difficult tasks to avoid "urged the speaker", as well as improve the Speaking Interval Detection (SID) module to detect murmur by the older adult participant and hence avoid interrupting their thinking process. Improvement of SID may also improve to RBC generation in a similar manner. We observed only one instance of inappropriate PBCs caused by the inability of the system to understand the speaker' utterance, which suggests that the use of Automatic Speech Recogntion (ASR) may not be essential for triggering PBCs.

Third, aside from reducing instances of inappropriate backchannel, we also need to improve participant-related adaptivity of backchanneling strategies. According to the Yerkes–Dodson law, each task performer has an optimal mental arousal to perform a certain task [90]. Clearly there exist individual differences among participants in terms of skill level (cognitive ability), personality and trait anxiety as potential influencers of arousal level [8, 23, 53]. As a result, the optimal arousal level may vary from person to person. In addition, individual difference in sensitivity may explain why some

participants did not notice RBCs while others did. We have attempted to integrate the idea of adaptive backchanneling based on participant characteristics encoded as *Participant Score*. This was a first step towards finding an optimal, personalized backchanneling strategy for each user. To develop such adaptive strategies, we will need to further conduct within-subject studies with multiple levels of proactivity in backchanneling. Other information about users e.g., language [14, 29], culture [17], etc. may also affect how users perceive backchannels, and thus may be incorporated into the personalized backchanneling strategy.

7.3 Generalization

Although our work was evaluated in a specific task scenario for a particular user population, it may be generalized in three ways. First, the backchanneling algorithm we propose may be applied to other kinds of task-oriented conversations requiring speech responses, including other cognitive tests [31], self-disclosure [43], counseling [35], to name a few. While the timing and frequency of backchanneling, especially for PBCs, may need to be adjusted according to the social rules and common human-human interaction practices in the target tasks, our data-driven approach to modeling backchannels has the potential to be easily adapted. Second, our CA system may be extended to serve other user groups, e.g., young adults who are vulnerable to neurocognitive disorders [67], children who need to be engaged in learning tasks [69], etc. Third, we demonstrated a workflow of investigating backchanneling patterns in Cantonese as a low-resource language. Different from well-studied languages like English, it is a challenge to study backchanneling for a language without previous work on this topic and supportive toolkits such as reliable ASR algorithms. Our work introduces potential solutions to overcome those constraints in both data analysis and implementation strategies, which might be beneficial for future research on backchanneling targeting low-resource languages.

7.4 Limitations and Future Work

There are several limitations regarding the design of the system and the experiment. As a proof-of-concept system, the TalkTive system is currently deployed on a desktop. We plan to develop a mobile APP version of it in the future, which offers accessibility via mobile phones. Also, the current study was conducted in the laboratory with a trial version of the MoCA test. The full version of the MCA test can be conducted with the mobile APP in the future. Another limitation associated with the in-lab setup is the difficulty in recruiting older adult participants due to their reduced mobility especially under the restrictions under COVID-19. Compared with the size of potential user population of the *TalkTive* system (i.e., the number of older adults in Hong Kong), the evaluation study was underpowered with 36 participants divided into three conditions. With more participants, our study will have a higher statistical power in analyzing the differences among various backchanneling strategies and revealing the general preference of older adults. With more participants, we will also be able to further analyze the relationship between the characteristics of the participants and their acceptance of backchanneling. This will help achieve greater personalization of backchanneling strategies. In addition, if a high-performance ASR system for older adult speech becomes available, the system may better understand the user's utterance and improve the trigger of RBCs.

Moreover, there still exist intrinsic limitations of CAs as compared with human healthcare professionals, such as not being as responsive and empathetic [37, 41, 86]. We believe that human therapists and CAs are complementary to each other and CAs should not aim to replace human healthcare professionals. In this study we have demonstrated that CAs present an affordable solution that enables older adults to take pre-screening tests in their homes and communities. For users considered to have a high risk of neurocognitive disorder based on pre-screening results, they need to seek further screening, diagnoses and treatment in the clinic. In sum, CAs have the potential in offering a scalable, accessible and economical means to support preliminary assessment of cognitive decline, and can help support the clinician's work to attain greater effectiveness and higher efficiency.

8 CONCLUSION

Conversational agents (CA) hold a strong promise in supporting digital cognitive assessments, with minimal human intervention, in order to scale up cognitive screening for early detection of NCDs. This paper presents an approach for automatic generation of backchanneling, a type of verbal response that enables the CA to acknowledge the user's input and encourages them to interact further. We analyzed a dataset with 246 human-human conversations involving an assessor and a participant in a the Montreal Cognitive Assessment (MoCA) test. We identified two kinds of backchannels – reactive backchannels (RBCs, e.g. "hmm") and proactive backchannels (PBCs, e.g. "what's more") – commonly adopted by human assessors. We labeled 2,732 RBCs and 2,037 PBCs in the MoCA dataset, and devised a data-driven method to model the timing of the two types of backchanneling. We proposed algorithms that generate RBCs and PBCs based on task-related and participant-related patterns, and developed TalkTive, a CA which can predict the timing and form of backchanneling while conducting speech-based cognitive assessments. For evaluation, we conducted a between-subject, in-laboratory study with n=36 older adult participants. The study demonstrated that the backchanneling algorithm can effectively generate PBCs and RBCs, over 88% of which were deemed appropriate by a human expert. Quantitative and qualitative evaluations in participant experience reflected that the automatically generated backchanneling was regarded as smoothly incorporated and not intrusive, while PBCs were preferred to RBCs.

ACKNOWLEDGMENTS

This project was supported by the HKSARG Research Grants Council's Theme-based Research Grant Scheme (Project No. T45-407/19N). We would like to thank Pauline Kwan for recording a Cantonese MoCA test and backchannel library, and Stephen MacNeil for feedback. We would also like to thank all the study participants for their time and feedback.

REFERENCES

[1] Sameera A Abdul-Kader and JC Woods. 2015. Survey on chatbot design techniques in speech conversation systems. *International Journal of Advanced Computer Science and Applications* 6, 7 (2015).

[2] Asangaedem Akpan, Maturin Tabue-Teguo, and Bertrand Fougère. 2019. Neurocognitive disorders: importance of early/timely detection in daily clinical practice. *Journal of Alzheimer's Disease* 70, 2 (2019), 317–322.

[3] Sabah Al-Hameed, Mohammed Benaissa, and Heidi Christensen. 2016. Simple and robust audio-based detection of biomarkers for Alzheimer's disease. In *7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*. 32–36.

[4] Alzheimer's Association. 2019. 2019 Alzheimer's disease facts and figures. *Alzheimer's & dementia* 15, 3 (2019), 321–387.

[5] Adrian Bangerter and Herbert H Clark. 2003. Navigating joint projects with dialogue. *Cognitive science* 27, 2 (2003), 195–225.

[6] Janet B Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of personality and social psychology* 79, 6 (2000), 941.

[7] Janet Beavin Bavelas and Jennifer Gerwing. 2011. The listener as addressee in face-to-face dialogue. *International Journal of Listening* 25, 3 (2011), 178–198.

[8] Daniel E Berlyne. 1960. Conflict, arousal, and curiosity. (1960).

[9] Roxane Bertrand, Gaëlle Ferré, Philippe Blache, Robert Espesser, and Stéphane Rauzy. 2007. Backchannels revisited from a multimodal perspective. In *Auditory-visual Speech Processing*. 1–5.

[10] Camiel J Beukeboom. 2009. When words feel right: How affective expressions of listeners change a speaker's language use. *European Journal of Social Psychology* 39, 5 (2009), 747–756.

[11] Harry N Boone and Deborah A Boone. 2012. Analyzing likert data. *Journal of extension* 50, 2 (2012), 1–5.

[12] Lawrence J Brunner. 1979. Smiles can be back channels. *Journal of Personality and Social Psychology* 37, 5 (1979), 728.

[13] Rupayan Chakraborty, Meghna Pandharipande, Chitralekha Bhat, and Sunil Kumar Kopparapu. 2020. Identification of dementia using audio biomarkers. *arXiv preprint arXiv:2002.12788* (2020).

[14] Patricia M Clancy, Sandra A Thompson, Ryoko Suzuki, and Hongyin Tao. 1996. The conversational use of reactive tokens in English, Japanese, and Mandarin. *Journal of pragmatics* 26, 3 (1996), 355–387.

[15] Herbert H Clark and Meredyth A Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of memory and language* 50, 1 (2004), 62–81.

[16] Mario Conci, Fabio Pianesi, and Massimo Zancanaro. 2009. Useful, social and enjoyable: Mobile phone adoption by older people. In *IFIP conference on human-computer interaction*. Springer, 63–76.

[17] Pino Cutrone. 2014. A cross-cultural examination of the backchannel behavior of Japanese and Americans: Considerations for Japanese EFL learners.

[18] Julie Doyle, Niamh Caprani, and Rodd Bond. 2015. Older adults' attitudes to self-management of health and wellness through smart home data. In *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. IEEE, 129–136.

[19] Julie Doyle, Lorcan Walsh, Antonella Sassu, and Teresa McDonagh. 2014. Designing a wellness self-management tool for older adults: results from a field trial of YourWellness. In *Proceedings of the 8th international conference on pervasive computing technologies for healthcare*. 134–141.

[20] Gary F. Simons Eberhard, David M. and Charles D. Fennig (eds.). 2021. Ethnologue: Languages of the World. (2021).

[21] Cleusa P Ferri, Martin Prince, Carol Brayne, Henry Brodaty, Laura Fratiglioni, Mary Ganguli, Kathleen Hall, Kazuo Hasegawa, Hugh Hendrie, Yueqin Huang, et al. 2005. Global prevalence of dementia: a Delphi consensus study. *The lancet* 366, 9503 (2005), 2112–2117.

[22] Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. 2016. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease* 49, 2 (2016), 407–422.

[23] Alinda Friedman. 1979. Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of experimental psychology: General* 108, 3 (1979), 316.

[24] Rod Gardner. 2001. *When listeners talk: Response tokens and listener stance*. Vol. 92. John Benjamins Publishing.

[25] Charles Goodwin. 1986. Between and within: Alternative sequential treatments of continuers and assessments. *Human studies* 9, 2 (1986), 205–217.

[26] Agustín Gravano and Julia Hirschberg. 2009. Backchannel-inviting cues in task-oriented dialogue. In *Tenth Annual Conference of the International Speech Communication Association*.

[27] Vicki L Hanson. 2010. Influencing technology adoption by older adults. *Interacting with Computers* 22, 6 (2010), 502–509.

[28] Kohei Hara, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara. 2018. Prediction of turn-taking using multitask learning with prediction of backchannels and fillers. *Listener* 162 (2018), 364.

[29] Bettina Heinz. 2003. Backchannel responses as strategic responses in bilingual speakers' conversations. *Journal of pragmatics* 35, 7 (2003), 1113–1142.

[30] Dirk Heylen, Elisabetta Bevacqua, Catherine Pelachaud, Isabella Poggi, Jonathan Gratch, and Marc Schröder. 2011. Generating listening behaviour. In *Emotion-oriented systems*. Springer, 321–347.

[31] S Hoops, S Nazem, AD Siderowf, JE Duda, SX Xie, MB Stern, and D Weintraub. 2009. Validity of the MoCA and MMSE in the detection of MCI and dementia in Parkinson disease. *Neurology* 73, 21 (2009), 1738–1745.

[32] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. 2010. Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*. 1265–1272.

[33] Shoichi Iwasaki. 1997. The Northridge earthquake conversations: The floor structure and the 'loop' sequence in Japanese conversation. *Journal of Pragmatics* 28, 6 (1997), 661–693.

[34] Vidit Jain, Maitree Leekha, Rajiv Ratn Shah, and Jainendra Shukla. 2021. Exploring Semi-Supervised Learning for Predicting Listener Backchannels. *arXiv preprint arXiv:2101.01899* (2021).

[35] Tatsuya Kawahara, Takashi Yamaguchi, Koji Inoue, Katsuya Takanashi, and Nigel G Ward. 2016. Prediction and Generation of Backchannel Form for Attentive Listening Systems.. In *Interspeech*. 2890–2894.

[36] Susan Kemper, Andrea Finter-Urczyk, Patrice Ferrell, Tamara Harden, and Catherine Billington. 1998. Using elderspeak with older adults. *Discourse Processes* 25, 1 (1998), 55–73.

[37] Junhan Kim, Sun Park, Lionel Robert, et al. 2019. Conversational agents for health and wellbeing: Review and future agendas. (2019).

[38] Alexandra König, Nicklas Linz, Johannes Tröger, Maria Wolters, Jan Alexandersson, and Phillipe Robert. 2018. Fully automatic speech-based analysis of the semantic verbal fluency task. *Dementia and geriatric cognitive disorders* 45, 3-4 (2018), 198–209.

[39] Anastasia Kononova, Lin Li, Kendra Kamp, Marie Bowen, RV Rikard, Shelia Cotten, and Wei Peng. 2019. The use of wearable activity trackers among older adults: focus group study of tracker perceptions, motivators, and barriers in the maintenance stage of behavior change. *JMIR mHealth and uHealth* 7, 4 (2019), e9832.

[40] Jarosław Kowalski, Anna Jaskulska, Kinga Skorupska, Katarzyna Abramczuk, Cezary Biele, Wiesław Kopeć, and Krzysztof Marasek. 2019. Older adults and voice interaction: A pilot study with google home. In *Extended Abstracts of the 2019 CHI Conference on human factors in computing systems*. 1–6.

[41] Kira Kretzschmar, Holly Tyroll, Gabriela Pavarini, Arianna Manzini, Ilina Singh, and NeurOx Young People's Advisory Group. 2019. Can your phone be your therapist? Young people's ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support. *Biomedical informatics insights* 11 (2019), 1178222619829083.

[42] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[43] Campbell Leaper, Mary Carson, Carilyn Baker, Heithre Holliday, and Sharon Myers. 1995. Self-disclosure and listener verbal support in same-gender and cross-gender friends' conversations. *Sex Roles* 33, 5-6 (1995), 387–404.

[44] Byung Cheol Lee, Junfei Xie, Toyin Ajisafe, and Sung-Hee Kim. 2020. How are wearable activity trackers adopted in older adults? Comparison between subjective adoption attitudes and physical activity performance. *International journal of environmental research and public health* 17, 10 (2020), 3461.

[45] Rock Leung, Charlotte Tang, Shathel Haddad, Joanna Mcgrenere, Peter Graf, and Vilia Ingriany. 2012. How older adults learn to use mobile devices: Survey and field investigations. *ACM Transactions on Accessible Computing (TACCESS)* 4, 3 (2012), 1–33.

[46] Joanna E Lewis and Mark B Neider. 2017. Designing wearable technology for an aging population. *Ergonomics in Design* 25, 3 (2017), 4–10.

[47] Jinchao Li, Jianwei Yu, Zi Ye, Simon Wong, Manwai Mak, Brian Mak, Xunying Liu, and Helen Meng. 2021. A comparative study of acoustic and linguistic features classification for alzheimer's disease detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6423–6427.

[48] Siân E Lindley, Richard Harper, and Abigail Sellen. 2009. Desiring to be in touch in a changing communications landscape: attitudes of older adults. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1693–1702.

[49] Gale M Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior* 37 (2014), 94–100.

[50] Tivadar Lucza, Kázmér Karádi, János Kállai, Rita Weintraut, József Janszky, Attila Makkos, Sámuel Komoly, and Norbert Kovács. 2015. Screening mild and major neurocognitive disorders in Parkinson's disease. *Behavioural neurology* 2015 (2015).

[51] Stephen MacNeil, Kyle Kiefer, Brian Thompson, Dev Takle, and Celine Latulipe. 2019. Ineqdetect: A visual analytics system to detect conversational inequality and support reflection during active learning. In *Proceedings of the ACM Conference on Global Computing Education*. 85–91.

[52] Xia Mao, Yiping Peng, Yuli Xue, Na Luo, and Alberto Rovetta. 2015. Backchannel Prediction for Mandarin Human-Computer Interaction. *IEICE TRANSACTIONS on Information and Systems* 98, 6 (2015), 1228–1237.

[53] Gerald Matthews, Timothy J Sparkes, and Helen M Bygrave. 1996. Attentional overload, stress, and simulate driving performance. *Human Performance* 9, 1 (1996), 77–101.

[54] Nicolai Meinshausen and Peter Bühlmann. 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72, 4 (2010), 417–473.

[55] Bahman Mirheidari, Daniel Blackburn, Traci Walker, Markus Reuber, and Heidi Christensen. 2019. Dementia detection using automatic analysis of conversations. *Computer Speech & Language* 53 (2019), 65–79.

[56] Bahman Mirheidari, Daniel Blackburn, Traci Walker, Annalena Venneri, Markus Reuber, and Heidi Christensen. 2018. Detecting Signs of Dementia Using Word Vector Representations.. In *Interspeech*. 1893–1897.

[57] Tracy L Mitzner, Julie B Boron, Cara Bailey Fausset, Anne E Adams, Neil Charness, Sara J Czaja, Katinka Dijkstra, Arthur D Fisk, Wendy A Rogers, and Joseph Sharit. 2010. Older adults talk technology: Technology usage and attitudes. *Computers in human behavior* 26, 6 (2010), 1710–1721.

[58] Louis-Philippe Morency, Iwan De Kok, and Jonathan Gratch. 2008. Predicting listener backchannels: A probabilistic multimodal approach. In *International Workshop on Intelligent Virtual Agents*. Springer, 176–190.

[59] Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. 2010. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems* 20, 1 (2010), 70–84.

[60] Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. 2005. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society* 53, 4 (2005), 695–699.

[61] Neal R Norrick. 2010. Incorporating recipient evaluations into stories. *Narrative Inquiry* 20, 1 (2010), 182–203.

[62] Neal R Norrick. 2010. Listening practices in television celebrity interviews. *Journal of Pragmatics* 42, 2 (2010), 525–543.

[63] Neal R Norrick. 2012. Listening practices in English conversation: The responses responses elicit. *Journal of Pragmatics* 44, 5 (2012), 566–576.

[64] United Nations Department of Economic and Social Affairs. 2019. World Population Ageing 2019: Highlights. (2019).

[65] Daniel Ortega, Chia-Yu Li, and Ngoc Thang Vu. 2020. Oh, Jeez! or uh-huh? A listener-aware Backchannel predictor on ASR transcriptions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8064–8068.

[66] Yilin Pan, Bahman Mirheidari, Markus Reuber, Annalena Venneri, Daniel Blackburn, and Heidi Christensen. 2019. Automatic Hierarchical Attention Neural Network for Detecting AD.. In *Interspeech*. 4105–4109.

[67] Peter K Panegyres and Kate Frencham. 2007. Course and causes of suspected dementia in young adults: a longitudinal study. *American Journal of Alzheimer's Disease & Other Dementias®* 22, 1 (2007), 48–56.

[68] Carolyn Pang, Zhiqin Collin Wang, Joanna McGrenere, Rock Leung, Jiamin Dai, and Karyn Moffatt. 2021. Technology Adoption and Learning Preferences for Older Adults: Evolving Perceptions, Ongoing Challenges, and Emerging Design Opportunities. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[69] Hae Won Park, Mirko Gelsomini, Jin Joo Lee, and Cynthia Breazeal. 2017. Telling stories to robots: The effect of backchanneling on a child's storytelling. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. IEEE, 100–108.

[70] Christina Patterson et al. 2018. World alzheimer report 2018. (2018).

[71] John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10, 3 (1999), 61–74.

[72] Anna Pompili, Alberto Abad, David Martins de Matos, and Isabel Pavão Martins. 2020. Pragmatic aspects of discourse production for the automatic identification of alzheimer's disease. *IEEE Journal of Selected Topics in Signal Processing* 14, 2 (2020), 261–271.

[73] Ronald Poppe, Khiet P Truong, and Dirk Heylen. 2011. Backchannels: Quantity, type and timing matters. In *International Workshop on Intelligent Virtual Agents*. Springer, 228–239.

[74] Ronald Poppe, Khiet P Truong, Dennis Reidsma, and Dirk Heylen. 2010. Backchannel strategies for artificial listeners. In *International Conference on Intelligent Virtual Agents*. Springer, 146–158.

[75] Alisha Pradhan, Leah Findlater, and Amanda Lazar. 2019. " Phantom Friend" or" Just a Box with Information" Personification and Ontological Categorization of Smart Speaker-based Voice Assistants by Older Adults. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.

[76] Dennis Reidsma, Iwan de Kok, Daniel Neiberg, Sathish Chandra Pammi, Bart van Straalen, Khiet Truong, and Herwin van Welbergen. 2011. Continuous interaction with a virtual human. *Journal on Multimodal User Interfaces* 4, 2 (2011), 97–118.

[77] Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2017. Enhancing Backchannel Prediction Using Word Embeddings.. In *INTERSPEECH*. 879–883.

[78] Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2019. Yeah, right, uh-huh: a deep learning backchannel predictor. In *Advanced Social Interaction with Agents*. Springer, 247–258.

[79] Emanuel A Schegloff. 1982. Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. *Analyzing discourse: Text and talk* 71 (1982), 93.

[80] Björn Schuller, Stefan Steidl, and Anton Batliner. 2009. The interspeech 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association*.

[81] Björn Schuller, Stefan Steidl, Anton Batliner, Simone Hantke, Florian Hönig, Juan Rafael Orozco-Arroyave, Elmar Nöth, Yue Zhang, and Felix Weninger. 2015. The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, Parkinson's & eating condition. In *Sixteenth annual conference of the international speech communication association*.

[82] Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, et al. 2012. The interspeech 2012 speaker trait challenge. In *Thirteenth annual conference of the international speech communication association*.

[83] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. 2013. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

[84] Korok Sengupta, Sayan Sarcar, Alisha Pradhan, Roisin McNaney, Sergio Sayago, Debaleena Chattopadhyay, and Anirudha Joshi. 2020. Challenges and Opportunities of Leveraging Intelligent Conversational Assistant to Improve the Well-being of Older Adults. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–4.

[85] Rajen D Shah and Richard J Samworth. 2013. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75, 1 (2013), 55–80.

[86] BG Silverman, N Hanrahan, L Huang, EF Rabinowitz, and S Lim. 2016. Artificial Intelligence in Behavioral and Mental Health Care.

[87] Thamar Solorio, Olac Fuentes, Nigel G Ward, and Yaffa Al Bayyari. 2006. Prosodic feature generation for back-channel prediction. In *Ninth International Conference on Spoken Language Processing*.

[88] Tanya Stivers. 2008. Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation. *Research on language and social interaction* 41, 1 (2008), 31–57.

[89] David F Tang-Wai, Eric E Smith, Marie-Andrée Bruneau, Amer M Burhan, Atri Chatterjee, Howard Chertkow, Samira Choudhury, Ehsan Dorri, Simon Ducharme, Corinne E Fischer, et al. 2020. CCCDTD5 recommendations on early and timely assessment of neurocognitive disorders using cognitive, behavioral, and functional scales. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* 6, 1 (2020), e12057.

[90] Karl Halvor Teigen. 1994. Yerkes-Dodson: A law for all seasons. *Theory & Psychology* 4, 4 (1994), 525–547.

[91] Jackson Tolins and Jean E Fox Tree. 2014. Addressee backchannels steer narrative development. *Journal of Pragmatics* 70 (2014), 152–164.

[92] Khiet P Truong, Ronald Poppe, and Dirk Heylen. 2010. A rule-based backchannel prediction model using pitch and pause information. In *Eleventh Annual Conference of the International Speech Communication Association*.

[93] Aditya Nrusimha Vaidyam, Hannah Wisniewski, John David Halamka, Matcheri S Kashavan, and John Blake Torous. 2019. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry* 64, 7 (2019), 456–464.

[94] John Vines, Gary Pritchard, Peter Wright, Patrick Olivier, and Katie Brittain. 2015. An age-old problem: Examining the discourses of ageing in HCI and strategies for future research. *ACM Transactions on Computer-Human Interaction (TOCHI)* 22, 1 (2015), 1–27.

[95] Nigel Ward. 1996. Using prosodic clues to decide when to produce back-channel utterances. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, Vol. 3. IEEE, 1728–1731.

[96] Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of pragmatics* 32, 8 (2000), 1177–1207.

[97] Jochen Weiner, Christian Herff, and Tanja Schultz. 2016. Speech-Based Detection of Alzheimer's Disease in Conversational German.. In *INTERSPEECH*. 1938–1942.

[98] Felix Weninger, Florian Eyben, Björn W Schuller, Marcello Mortillaro, and Klaus R Scherer. 2013. On the acoustics of emotion in audio: what speech, music, and sound have in common. *Frontiers in psychology* 4 (2013), 292.

[99] Bo Xiao, Panayiotis G Georgiou, Zac E Imel, David C Atkins, and Shrikanth S Narayanan. 2013. Modeling therapist empathy and vocal entrainment in drug addiction counseling.. In *INTERSPEECH*. 2861–2865.

[100] Zi Ye, Shoukang Hu, Jinchao Li, Xurong Xie, Mengzhe Geng, Jianwei Yu, Junhao Xu, Boyang Xue, Shansong Liu, Xunying Liu, et al. 2021. Development of the cuhk elderly speech recognition system for neurocognitive disorder detection using the dementiabank corpus. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6433–6437.

[101] Robert M Yerkes, John D Dodson, et al. 1908. The relation of strength of stimulus to rapidity of habit-formation. *Punishment: Issues and experiments* (1908), 27–41.

[102] Victor H Yngve. 1970. On getting a word in edgewise. In *Chicago Linguistics Society, 6th Meeting, 1970*. 567–578.

[103] Richard F Young and Jina Lee. 2004. Identifying units in interaction: Reactive tokens in Korean and English conversations. *Journal of Sociolinguistics* 8, 3 (2004), 380–407.

[104] Ruby Yu, PH Chau, SM McGhee, WL Cheung, KC Chan, SH Cheung, and J Woo. 2010. Dementia trends: Impact of the ageing population and societal implications for Hong Kong.

[105] Randall Ziman and Greg Walsh. 2018. Factors affecting seniors' perceptions of voice-enabled user interfaces. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.

[106] Tamara Zubatiy, Kayci L Vickers, Niharika Mathur, and Elizabeth D Mynatt. 2021. Empowering Dyads of Older Adults With Mild Cognitive Impairment And Their Care Partners Using Conversational Agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.