Automatic mild cognitive impairment estimation from the group conversation of coimagination method

Sixia Li
Japan Advanced Institute of Science and Technology
Japan
lisixia@jaist.ac.jp

Kazumi Kumagai
Center for Advanced Intelligence Project, RIKEN
Japan
kazumi.kumagai@riken.jp

Mihoko Otake-Matsuura
Center for Advanced Intelligence Project, RIKEN
Japan
mihoko.otake@riken.jp

Shogo Okada*
Japan Advanced Institute of Science and Technology
Japan
okada-s@jaist.ac.jp

Abstract

The coimagination method (Otake, 2009) is designed to prevent dementia in individuals with mild cognitive impairment (MCI) by utilizing the brain's natural processes. This method involves participants sharing their thoughts and feelings through group conversations centered around shared photos. The coimagination method contains two phases: (1) each participant talk about their memories and experiences related to the photos they bring, and (2) other participants ask questions about the photos. Automating the MCI estimation could be helpful for assisting individuals with MCI during coimagination. However, previous MCI estimation methods rarely focused on group conversation scenarios, despite the potential of multimodal features observed in these scenarios in revealing cognitive states. This study focuses MCI individuals defined by cognitive test scores (e.g., Mini-Mental State Examination (MMSE)). We explore MCI estimation from three aspects. First, we clarify whether MCI can be effectively estimated by constructing estimation models using linguistic and acoustic features from coimagination sessions. Second, we evaluate the impact of using data from the two distinct phases, as they may activate participants' cognitive functions differently. Finally, we analyze the effects of incorporating subtasks including participants' conversational customary and engagement level during coimagination via multitask learning. The experimental results demonstrated that individuals with MCI can be effectively estimated from group conversations from coimagination, with the highest macro F1 score of 0.693. The results also demonstrated that the performance improved when using data from the phase that highly activates cognitive functions and when considering conversation customary as a subtask.

CCS Concepts

• **Human-centered computing** → *Empirical studies in ubiquitous and mobile computing*; *Human computer interaction (HCI).*

*Shogo Okada is the corresponding author

ICMI '24, November 04–08, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0462-8/24/11
https://doi.org/10.1145/3678957.3685754

Keywords

Mild cognitive impairment estimation, coimagination method, group conversation, machine learning

ACM Reference Format:
Sixia Li, Kazumi Kumagai, Mihoko Otake-Matsuura, and Shogo Okada. 2024. Automatic mild cognitive impairment estimation from the group conversation of coimagination method. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 04–08, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3678957.3685754

1 Introduction

Dementia is a crucial and increasing global problem [12], but early intervention holds potential for prevention [19]. Coimagination is an effective method to prevent dementia [21] based on the hypothesis that activating cognitive functions, which begin to decline at the mild cognitive impairment (MCI) stage, is effective [3, 26].

Coimagination is a conversation-based method that involves having participants share their thoughts and feelings for a set amount of time through the use of pictures and related topics. One coimagination session consists of two phases with four participants and one host [24]. The left part of Figure 1 shows an image of the two phases. Each participant prepares a fixed number of photos (two in this study) in advance. In the first phase for introduction, every participant introduces their photos in order, speaking for one minute per photo. In the second phase for question-answering, each participant is questioned by the other participants. Each other participant must ask at least one question to the participant currently being questioned. The host observes participants' actions; if one participant does not ask any questions, the host will ask the participant to ask question. Question-answering process for each photo is limited to two minutes. Coimagination has been shown to be effective at activating MCI individuals' cognitive functions [21, 24] and is thus implemented in many places [22, 23].

With the popularization of coimagination, automatically estimating MCI individuals could be helpful in assisting these individuals during coimagination. However, MCI estimation is a challenging task. Most previous studies focused on distinguishing MCI individuals from others [25] and mainly focused on the scenario of participant speaking alone [5] or during general daily activities [18, 20, 28]. Few studies focused on estimating MCI from multimodal behaviors in group conversations, such as those in coimagination. In addition

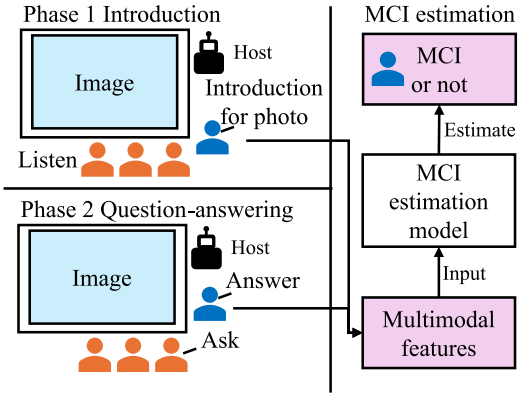

Figure 1: MCI estimation from group conversation of coimagination. The left part shows the two phases of coimagination, and the right part shows the MCI estimation process.

to participants' multimodal features, interactions among participants are important cues in capturing cognitive states in group conversations, further increasing the complexity of MCI estimation.

For this challenging task, we explore three aspects of MCI estimation from group conversations during coimagination sessions:

1. We construct MCI estimation models using linguistic and acoustic features from coimagination group conversations. The experimental results demonstrated that interaction ability-related and cognitive-related features are effective for MCI estimation.

2. Since participants' cognitive functions may differ in the two phases of coimagination and are thus reflected by different multimodal signals, we examine the effect of using data from each phase. We found that using data from the question-answering phase, where participants need to activate cognitive functions more, led to higher macro F1 score than using the introduction phase data.

3. We investigate the effect of considering subtasks for MCI estimation, as these subtasks can influence behaviors in conversations, and consequently, the relationship between multimodal behaviors and MCI. The results showed that considering conversation customary as a subtask improved the MCI estimation performance.

2 Dataset

In this study, we use a Japanese coimagination dataset collected by RIKEN in 2018 [24]. The dataset excluded participants with a Mini-Mental State Examination (MMSE) score ≤ 23 that are clearly dementia. The dataset comprises nine groups, with four participants in each group. Among these participants, 3-4 were permanent participants who regularly participated, and 0-1 were temporary participants that are typically filled by healthy staff members. Each group conducted 12 sessions. Permanent participants underwent the MMSE and Montreal Cognitive Assessment (MoCA) tests twice: before the first session and after the last session. Permanent participants also annotated their conversation customary and engagement levels in coimagination. We treat participants in each session as independent participants to increase data samples.

MCI labels are annotated based on the MMSE and MoCA thresholds. The MoCA threshold is generally defined as 25 [8]. A MoCA score of ≤ 25 is considered MCI, and a score of 26-30 is considered healthy. Various MMSE thresholds exist. We use a threshold of 26, as it is commonly used to define the clinical spectrum of Alzheimer's disease [1, 6]. Accordingly, an MMSE score of ≤ 26 is considered MCI, and a score of 27-30 is considered healthy. In annotating, if a participant's score is under either the MMSE or MoCA threshold at least once, we label the participant as MCI; otherwise, healthy.

Table 1: Statistics of MCI categories and conversation-related characteristic categories in each MCI category after data cleaning

Category	Participants	Conversation customary (low/high)	Engagement level in coimagination (low/high)
Healthy	221	121/100	131/90
MCI	97	64/33	31/66

Conversation customary describes conversation customary in daily life, which is related to cognitive situation. We use a 4-point Likert scale to measure 12 items for each participant, such as "When communicating with people, do you understand them well and actively ask questions to them?" The lower a item score is, the more the participant tends to have that customary in daily life. We sum the scores of all the items of each participant and find the median value. If a participant's score is equal to or lower than the median, we label the participant as category low; otherwise, high.

Engagement level in coimagination measures overall how participants felt engaged in coimagination, which is an affective aspect that could be related to social behaviors and cognitive states [16]. We use a 3-point Likert scale to measure five engagement items, such as "Did you enjoy the coimagination?" The higher the score is, the more likely the participant engaged in coimagination. We sum the scores of all the items of each participant and find the median value. If a participant's score is equal to or lower than the median, we label the participant as category low; otherwise, high.

Table 1 lists the data statistics after data cleaning.

3 MCI estimation

This study aims to construct models to estimate MCI participants from their multimodal behaviors in group conversations of coimagination. Accordingly, this tasks is a binary classification task.

3.1 Multimodal features and DNN model

For multimodal features, we consider that participants' cognitive state may not be reflected in short conversations. Therefore, we use the mean feature among the whole conversation for each participant. We do not use time sequences of utterances because the data samples are relatively small, and modeling time sequences that usually use models with more parameters leads to overfitting.

Linguistic features Linguistic features can reflect participants' ability to think and organize language. We use five linguistic features that could be related to cognitive situations in this study.

Bidirectional encoder representations from transformers (BERT) BERT is effective for semantic representation [13, 15]. We use the cl-tohoku/bert-base-japanese model to obtain representations for each utterance.

BERT-Diff Cognitive states can influence linguistic behavior in adjacent utterances. To represent the linguistic behavior of the participant who took the turn, we compute the difference in BERT embedding between two participants' adjacent utterances.

Number of spoken utterances (N-utt) N-utt reflects how active a participant was. We count each complete sentences as an utterance. If a participant continuously spoke multiple complete sentences, each was counted as a separate utterance.

Part-Of-Speech (POS) The POS is shown to be related to one's internal state [30]. We use the Fugashi [17] toolkit to segment each utterance into words with POS parsing. Then we count POS to create bag-of-words vectors as the feature.

Linguistic inquiry and word count (LIWC) LIWC [29] describes word functions (e.g., firstly: function, conjunction). We apply the Japanese LIWC dictionary [11] to obtain word functions for each word in each utterance. We use the 'others' function for words that are not included in the LIWC dictionary. For each utterance, we create a bag-of-words vector by counting functions of all words.

Acoustic features Acoustic features are related to cognitive situations [4]. We use three acoustic features in this study.

InterSpeech2009 (IS09) IS09 is a feature set that was originally proposed for emotion recognition [27]. It describes high-level functionals of low-level descriptors (LLDs) including f0 and MFCC. We use OpenSmile [10] to extract IS09 for each utterance.

Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) eGeMAPS is a feature set that aims to provide a basic standard acoustic parameter set for automatic voice analysis [9]. Notably, eGeMAPS contains formants features that are sensitive to mental states and are effective in cognitive load classification tasks [9, 31]. We use OpenSmile to extract eGeMAPS for each utterance.

Wav2Vec2.0 (W2V) Wav2Vec2.0 [2] is a large-scale pretrained model that is effective for audio representation [7, 14]. We use the jonatasgrosman/wav2vec2-large-xlsr-53-japanese model to extract embeddings for each utterance and use the average pooling of the last layer [7] to obtain the features.

DNN model In this study, we use feedforward deep neural network (DNN) model that is widely used in various tasks, including MCI-related tasks [25]. We use a DNN model with three layers, each with 256 units. We use early fusion to concatenate the multimodal features into one vector as the input. After the three layers, a projection layer is used to obtain classification results. In the multitask learning, two projection layers are used for each task.

3.2 Phase separation

The activating of cognitive functions probably differs in the two phases of the coimagination method. In the introduction phase, participants generally only need to talk or listen. Whereas in the question-answering phase, participants need to think frequently to ask and answer questions. Cognitive states can be reflected differently via multimodal behaviors in the two phases. We use data from each phase to train models and then investigate the effect of using data from different types of cognitive function activation.

3.3 Multitask learning

Conversation-related characteristics can influence behavior in conversations, thus they may influence how multimodal behaviors reflect MCI. Table 1 shows that the distributions of these characteristics differ between individuals with MCI and healthy controls. Accordingly, these characteristics could be helpful for estimating MCI from multimodal features. To evaluate the impact of considering these characteristics for MCI estimation, we use multitask learning to simultaneously learn shared information between the MCI estimation and two specific characteristics as subtasks.

4 Experimental settings

We train the models via 5-fold cross validation and speaker independent settings. In particular, we split participant data into five folds and train five models separately, setting each fold for testing and the remaining folds for training. We use 90% of the data for training and 10% for validation. The participants in the training set and test set were guaranteed to be different to avoid information leakage. The average performance among the five models is used for evaluation. We use the macro F1 score as an evaluation metric.

In the experiment, we use all combinations (239) of modalities. For features other than BERT, BERT-Diff, and W2V, we use the z score to normalize the training and test sets separately. We set the maximum training epoch to 200 and do not stop training in the first 50 epochs to warm up. We use early stopping if the validation performance does not improve within five epochs after the 50th epoch. We use cross-entropy loss functions and the Adam optimizer with a learning rate of 0.0005. For multitask learning, we use a weighted sum with a weight of 0.8 for the main task and a weight of 0.2 for the subtask to compute the loss. We run all the experiments three times and use the average performance for evaluation to reduce the influence of random initialization.

5 Results and discussion

Tables 2 and 3 show the results of different tasks and phase separations. We only show the modality combinations that yield the best performance in each task setting. The unimodal performance are provided in Appendix A. The use of multimodal features improved the accuracy compared to that using unimodal features alone.

In Tables 2 and 3, the modality function and modality columns list the modality combinations with each modality contributing specific information. For example, the third row of Table 2 shows the results of the BERT-Diff + eGeMAPS combination. BERT-Diff reflects the characteristics of turn-taking, and eGeMAPS contains cognitive-related acoustic features. Therefore, these modalities are listed under turn-taking and cognitive-related acoustics, respectively. The bold and underlined numbers indicate the best performance of the corresponding task. w/customary and w/engagement indicate multitask learning via conversation customary and engagement in coimagination as subtasks, respectively.

MCI estimation results As shown in Table 2, the best model for all phases achieved a macro F1 score of 0.693. This result is far better than the random classification performance, which is 0.5. Moreover, a recent relevant work that also used linguistic and acoustic features to distinguish MCI and dementia individuals achieved macro F1 scores of 0.745 [5], although their dataset differs from ours. Accordingly, our results can be considered good. Therefore, the results demonstrated that MCI can be effectively estimated from multimodal behaviors in group conversations of coimagination.

Table 2: Macro F1 scores achieved using the best modality combinations for different tasks with data from all phases

Modality Functions and Modality				Single task	Multitask	
Semantic	Turn-taking	General acoustic	Cognitive-related acoustic	MCI	w/ customary	w/ engagement
BERT	BERT-Diff	W2V	eGeMAPS	0.677 (±0.018)	**0.707** (±0.018)	**0.699** (±0.006)
	BERT-Diff		eGeMAPS	**0.693** (±0.013)	0.681 (±0.006)	0.690 (±0.017)

Table 3: Macro F1 scores achieved using the best modality combinations for different tasks and phase separation. (a) Results of using data from the introduction phase only; (b) results of using data from the question-answering phase only

(a) Best macro F1 scores using data averaged from the introduction phase

Modality Functions and Modality					Single task	Multitask	
Semantic	Active	Turn-taking	Word function	Cognitive-related acoustic	MCI	w/ customary	w/ engagement
BERT	N-utt	BERT-Diff	LIWC	eGeMAPS	**0.655** (±0.011)	0.618 (±0.025)	0.640 (±0.018)
BERT		BERT-Diff	LIWC	eGeMAPS	0.650 (±0.030)	0.649 (±0.040)	**0.645** (±0.032)
	N-utt	BERT-Diff	LIWC	eGeMAPS	0.646 (±0.004)	**0.657** (±0.017)	0.630 (±0.025)

(b) Best macro F1 scores using data averaged from the question-answering phase

Modality Functions and Modality				Single task	Multitask	
Semantic	Active	Turn-taking	Cognitive-related acoustic	MCI	w/ customary	w/ engagement
BERT	N-utt	BERT-Diff	eGeMAPS	**0.693** (±0.008)	0.689 (±0.003)	**0.686** (±0.024)
		BERT-Diff	eGeMAPS	0.645 (±0.051)	**0.696** (±0.016)	0.651 (±0.010)

Effect of using different phases By comparing the best performance for a single task in Table 3 (a) and (b), it is observed that the use of data from the question-answering phase outperformed that using data from the introduction phase by 0.038 in terms of the F1 score. Since the question-answering phase activates cognitive functions more in terms of thinking and responding, the results demonstrated that multimodal behaviors in the high cognitive function activation phase could better reflect cognitive states.

Effect of using conversation-related characteristics By comparing the single task and multitask results in Tables 2 and 3, the best performance was achieved when using conversation customary as the subtask, with improvements in the best performance over that of single tasks in all phase settings. On the other hand, the best performance achieved when using engagement only improved the best performance of single task when data from all phases was used. We speculate that the reason is that conversation customary reflects language organization abilities in daily life, which are highly related to cognitive functions. However, engagement is more related to feelings, not to cognition. Therefore, considering conversation customary as a subtask could assist in estimating MCI, while considering engagement rarely learn cognitive-related shared information. These results suggest that considering cognitive-related tasks as subtasks is effective for MCI estimation.

Effective modalities As shown in Tables 2 and Table 3 (b), the best modality combination when using data from all phases and from the question-answering phase is BERT-Diff + eGeMAPS. BERT-Diff reflects the characteristic of how a participant organizes language in adjacency utterances to respond to the previous turn in an interaction. eGeMAPS contains formant features that are effective in cognitive-related tasks [9, 31]. Therefore, the results demonstrated that considering features that describe characteristics in turn-taking and cognitive state-related features is effective in estimating MCI from group conversations.

As shown in Table 3 (a), in addition to BERT-Diff and eGeMAPS, LIWC is included in all the combinations with the best performance when the introduction phase data are used. In the introduction phase, participants are required to organize their language to tell a story related to a photo; thus, arranging words that serve suitable functions is necessary. Such arrangements could be influenced by one's cognitive state. Therefore, these results demonstrated that using features that describe word functions is effective in estimating MCI using data from the introduction phase in group conversation.

6 Conclusion

In this study, for the automatic estimation of MCI individuals and assisting these individuals during coimagination, we explored whether MCI can be effectively estimated from multimodal behaviors in group conversations of coimagination from three aspects. The experimental results demonstrated that MCI can be effectively estimated from multimodal behaviors. Furthermore, using data from the phase that acitivates cognitive function more is effective. In addition, using multitask learning with cognitive-related subtasks improves the MCI estimation performance. The results also suggested that features related to communication ability and the cognitive state are generally effective for MCI estimation and that features that describe word functions are effective for estimating MCI from the introduction phase of group conversations. In future works, designing specific model structures and features could be helpful for improving MCI estimation from group conversations.

Acknowledgments

This work was also partially supported by JSPS KAKENHI (22K21304, 22H04860, 22H00536, 23H03506), JST AIP Trilateral AI Research, Japan (JPMJCR20G6) and JST Moonshot R&D program (JPMJMS2237).

References

[1] Ingrid Arevalo-Rodriguez, Nadja Smailagic, Marta Roqué i Figuls, Agustín Ciapponi, Erick Sanchez-Perez, Antri Giannakou, Olga L Pedraza, Xavier Bonfill Cosp, and Sarah Cullum. 2015. Mini-Mental State Examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI). *Cochrane database of systematic reviews* 3 (2015).

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33 (2020), 12449–12460.

[3] Pascale Barberger-Gateau, Colette Fabrigoule, Isabelle Rouch, Luc Letenneur, and Jean-Francois Dartigues. 1999. Neuropsychological correlates of self-reported performance in instrumental activities of daily living and prediction of dementia. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 54, 5 (1999), P293–P303.

[4] Flavio Bertini, Davide Allevi, Gianluca Lutero, Danilo Montesi, and Laura Calzà. 2021. Automatic speech classifier for mild cognitive impairment and early dementia. *ACM Transactions on Computing for Healthcare (HEALTH)* 3, 1 (2021), 1–11.

[5] Laura Calzà, Gloria Gagliardi, Rema Rossini Favretti, and Fabio Tamburini. 2021. Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. *Computer Speech & Language* 65 (2021), 101113.

[6] Kimberly R Chapman, Hanaan Bing-Canar, Michael L Alosco, Eric G Steinberg, Brett Martin, Christine Chaisson, Neil Kowall, Yorghos Tripodis, and Robert A Stern. 2016. Mini Mental State Examination and Logical Memory scores for entry into Alzheimer's disease trials. *Alzheimer's research & therapy* 8 (2016), 1–11.

[7] Li-Wei Chen and Alexander Rudnicky. 2023. Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[8] Anne M Damian, Sandra A Jacobson, Joseph G Hentz, Christine M Belden, Holly A Shill, Marwan N Sabbagh, John N Caviness, and Charles H Adler. 2011. The Montreal Cognitive Assessment and the mini-mental state examination as screening instruments for cognitive impairment: item analyses and threshold scores. *Dementia and geriatric cognitive disorders* 31, 2 (2011), 126–131.

[9] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing* 7, 2 (2015), 190–202.

[10] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.

[11] Tasuku Igarashi, Shimpei Okuda, and Kazutoshi Sasahara. 2022. Development of the japanese version of the linguistic inquiry and word count dictionary 2015. *Frontiers in psychology* 13 (2022), 841534.

[12] Ashir Javeed, Ana Luiza Dallora, Johan Sanmartin Berglund, Arif Ali, Liaqata Ali, and Peter Anderberg. 2023. Machine learning for dementia prediction: a systematic review and future research directions. *Journal of medical systems* 47, 1 (2023), 17.

[13] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.

[14] Seongbin Kim, Gyuwan Kim, Seongjin Shin, and Sangmin Lee. 2021. Two-stage textual knowledge distillation for end-to-end spoken language understanding. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7463–7467.

[15] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.

[16] Leena Mathur, Maja Mataric, and Louis-Philippe Morency. 2023. Expanding the Role of Affective Phenomena in Multimodal Interaction Research. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 253–260.

[17] Paul McCann. 2020. fugashi, a Tool for Tokenizing Japanese in Python. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*. Association for Computational Linguistics, Online, 44–51. https://www.aclweb.org/anthology/2020.nlposs-1.7

[18] Akira Minamisawa, Shogo Okada, Ken Inoue, and Mami Noguchi. 2022. Dementia scale score classification based on daily activities using multiple sensors. *IEEE Access* 10 (2022), 38931–38943.

[19] Sandra Morovic, Hrvoje Budincevic, Valbona Govori, and Vida Demarin. 2019. Possibilities of dementia prevention-it is never too early to start. *Journal of medicine and life* 12, 4 (2019), 332.

[20] Shogo Okada, Ken Inoue, Toru Imai, Mami Noguchi, and Kaiko Kuwamura. 2019. Dementia scale classification based on ubiquitous daily activity and interaction sensing. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 192–198.

[21] Mihoko Otake, Motoichiro Kato, Toshihisa Takagi, and Hajime Asama. 2009. Coimagination method: Communication support system with collected images and its evaluation via memory task. In *Universal Access in Human-Computer Interaction. Addressing Diversity: 5th International Conference, UAHCI 2009, Held as Part of HCI International 2009, San Diego, CA, USA, July 19-24, 2009. Proceedings, Part I 5*. Springer, 403–411.

[22] Mihoko Otake, Motoichiro Kato, Toshihisa Takagi, Shuichi Iwata, Hajime Asama, and Jun Ota. 2011. Multiscale service design method and its application to sustainable service for prevention and recovery from dementia. In *New Frontiers in Artificial Intelligence: JSAI-isAI 2010 Workshops, LENLS, JURISIN, AMBN, ISS, Tokyo, Japan, November 18-19, 2010, Revised Selected Papers 2*. Springer, 321–330.

[23] Mihoko Otake-Matsuura, Yoshie Taguchi, Katsutoshi Negishi, Mitsuteru Matsumura, Kiyomi Shimizu, Eiko Nagata, Hideko Nagahisa, Akane Uotani, Akira Suzuki, Mieko Yoshida, et al. 2020. Services for Cognitive Health Co-created with Older Adults. In *Human Aspects of IT for the Aged Population. Technologies, Design and User Experience: 6th International Conference, ITAP 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I 22*. Springer, 59–72.

[24] Mihoko Otake-Matsuura, Seiki Tokunaga, Kumi Watanabe, Masato S Abe, Takuya Sekiguchi, Hikaru Sugimoto, Taishiro Kishimoto, and Takashi Kudo. 2021. Cognitive intervention through photo-integrated conversation moderated by robots (PICMOR) program: a randomized controlled trial. *Frontiers in Robotics and AI* 8 (2021), 633076.

[25] Chathurika Palliya Guruge, Sharon Oviatt, Pari Delir Haghighi, and Elizabeth Pritchard. 2021. Advances in multimodal behavioral analytics for early dementia diagnosis: A review. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 328–340.

[26] Dorene M Rentz and Sandra Weintraub. 2000. Neuropsychological detection of early probable Alzheimer's disease. In *Early diagnosis of Alzheimer's disease*. Springer, 169–189.

[27] B Schuller, S Steidl, and A Batliner. 2009. The Interspeech 2009 Emotion Challenge. In *Proc. Interspeech 2009, Brighton, UK*. 312–315.

[28] Shinichi Sugiura, Shinichiro Yokoyama, Ken Inoue, and Shogo Okada. 2023. Dementia Scale Classification with Sequential Model from Sleep Activity Data. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 1–5.

[29] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[30] Wenqing Wei, Sixia Li, and Shogo Okada. 2022. Investigating the relationship between dialogue and exchange-level impression. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 359–367.

[31] Tet Fei Yap, Julien Epps, Eliathamby Ambikairajah, and Eric HC Choi. 2011. Formant frequencies under cognitive load: Effects and classification. *EURASIP journal on advances in signal processing* 2011 (2011), 1–11.

A Appendix

A.1 Results of unimodality

Table A1 lists the performance of each unimodality in single -task and multitask settings. As shown in this table, the best result for single tasks is achieved using IS09. For multitask learning, W2V and eGeMAPS achieved the best performance when using conservation customary and engagement in coimagination subtasks, respectively.

Table A1: Unimodality results for each task using data from all phases

	Single	Multitask	
Modality	MCI	w/ customary	w/ engagement
BERT	0.549	0.532	0.538
BERT-Diff	0.558	0.574	0.567
N-utt	0.415	0.415	0.415
POS	0.486	0.459	0.477
LIWC	0.523	0.502	0.533
IS09	**0.628**	0.562	0.622
eGeMAPS	0.600	0.614	**0.637**
W2V	0.624	**0.629**	0.635

Table A2: Unimodality results for different tasks and phase separation. (a) Results of using data from the introduction phase only; (b) results of using data from the question-answering phase only

(a) Results using data from the introduction phase

	Single	Multitask	
Modality	MCI	w/customary	w/engagement
BERT	0.433	0.431	0.444
BERT-Diff	0.515	0.474	0.521
N-utt	0.434	0.442	0.458
POS	0.428	0.459	0.427
LIWC	0.488	0.462	0.511
IS09	**0.604**	0.547	0.568
eGeMAPS	0.590	**0.569**	**0.587**
W2V	0.553	0.553	0.581

(b) Results using data from the question-answering phase

	Single	Multitask	
Modality	MCI	w/customary	w/engagement
BERT	0.593	0.576	0.597
BERT-Diff	0.552	0.576	0.549
N-utt	0.415	0.415	0.415
POS	0.455	0.459	0.452
LIWC	0.537	0.521	0.537
IS09	0.603	0.597	0.589
eGeMAPS	**0.606**	0.626	0.585
W2V	0.599	**0.627**	**0.603**

Table A2 lists the performance of each unimodality in each task setting and phase separation. As shown in the table, the best performance is achieved by using acoustic features. These results are consistent with those in Table A1, suggesting that when focusing on a single modality, acoustic features are more effective for MCI estimation from group conversation.

On the other hand, compared with the results in Table 2 and Table 3 in the results and discussion section, the best performance using multimodal features outperformed the best performance using unimodality by an average of 0.07. The comparison results suggest that multimodal behaviors are more effective than unimodality behaviors for MCI estimation from group conversations.

A.2 Visualization of representations before the projection layer

We show what different models learned by using single task and multitask via t-distributed stochastic neighbor embedding (t-SNE) visualizations. Figure A1 shows 2-dimensional t-SNE visualizations of the middle representations before the projection layer of the DNN model when using best modality combination BERT+BERT-Diff+W2V+eGeMAPS. In Figure A1, the blue, green, and red symbols indicate representations using single task, multitask with conversations customary, and multitask with engagement in coimagination, respectively. The "x" marker indicates the representations of the healthy category, and the "▲" marker indicates the representations of the MCI category.

As shown in Figure A1, representations of using single task and multitask are distributed in different spaces; this visualization demonstrates that considering multitasks in estimating MCI influences the learning results. On the other hand, the healthy and MCI categories did not significantly differ across all three task settings. These results are consistent with the results in Table 2 that the use of multitask learning did not significantly improve the performance of using single task.

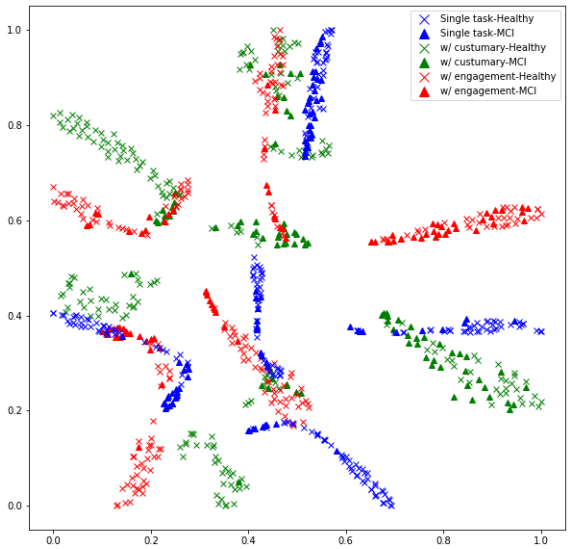

Figure A1: t-SNE visualization of middle representations using single task and multitask learning