Received: 27 August 2023 | Revised: 4 December 2023 | Accepted: 20 December 2023

DOI: 10.4218/etrij.2023-0356

SPECIAL ISSUE

ETRI Journal WILEY

Alzheimer's disease recognition from spontaneous speech using large language models

Jeong-Uk Bang[1] | **Seung-Hoon Han**[1,2] | **Byung-Ok Kang**[1]

[1]Integrated Intelligence Research Section, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea

[2]Department of Brain & Cognitive Engineering, Korea University, Seoul, Republic of Korea

Correspondence
Jeong-Uk Bang, Integrated Intelligence Research Section, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea.
Email: jubang0219@etri.re.kr

Funding information
This research was supported by the National Research Council of Science & Technology (NST) grant by the Korea government (MSIT) (No. CAP21052 300)

Abstract

We propose a method to automatically predict Alzheimer's disease from speech data using the ChatGPT large language model. Alzheimer's disease patients often exhibit distinctive characteristics when describing images, such as difficulties in recalling words, grammar errors, repetitive language, and incoherent narratives. For prediction, we initially employ a speech recognition system to transcribe participants' speech into text. We then gather opinions by inputting the transcribed text into ChatGPT as well as a prompt designed to solicit fluency evaluations. Subsequently, we extract embeddings from the speech, text, and opinions by the pretrained models. Finally, we use a classifier consisting of transformer blocks and linear layers to identify participants with this type of dementia. Experiments are conducted using the extensively used ADReSSo dataset. The results yield a maximum accuracy of 87.3% when speech, text, and opinions are used in conjunction. This finding suggests the potential of leveraging evaluation feedback from language models to address challenges in Alzheimer's disease recognition.

KEYWORDS

Alzheimer's disease, dementia, dementia detection, large language model, pretrained model

1 | INTRODUCTION

Alzheimer's disease (AD) is a chronic neurodegenerative disease that affects memory, thinking, cognitive skills, and the ability to perform simple tasks [1]. Given its prevalence, dementia has emerged as a critical global issue that must be addressed, owing to its substantial socioeconomic impact. Due to its severity, institutions and researchers worldwide are investing heavily in dementia prevention and early detection.

Typically, AD screening involves cognitive tests administered using pens and paper, such as the Mini-Mental State Examination (MMSE) [2] and Montreal Cognitive Assessment [3]. This scoring process relies on the subjective judgment of the clinical practitioner, which potentially results in errors and inter-rater variability [4]. To address these issues, numerous researchers [5–8] have focused on the development of automated assessments.

One promising direction is the detection of cognitive impairment based on speech signals [8]. Speech signals offer the advantage of being collected naturally and continuously throughout the day, thus allowing the accumulation of a significant amount of data without imposing a burden on the participants [9]. Furthermore, advancements in artificial intelligence and machine learning have led to substantial improvements in speech analysis over the past decade [10]. Recent research indicates the potential of speech as a biomarker of AD [5, 6, 11–13].

The ADReSSo Challenge [12] aimed to detect cognitive impairment and decline using spontaneous speech audio. In this challenge, participants were presented with the "Cookie Theft" picture from the Boston Diagnostic Aphasia Examination [14] and asked to describe everything they saw in the image. Their descriptions were recorded and used as benchmark datasets for AD detection. This challenge systematically compared various approaches [15–22] that focused on audio signals and speech recognition outputs (without manual transcription). In [15–18] pretrained models were used, such as Wav2Vec 2.0 [23] and Bidirectional Encoder Representations from Transformers (BERT) [24], achieving high performance. Additionally, studies using diverse information, such as handcrafted features [9, 12], speaker information [19, 20], emotion cues [21], and automatic speech recognition (ASR) confidence scores [22], have been presented.

We propose an approach for AD detection using the ChatGPT large language model (LLM) [25]. Patients with AD exhibit distinct characteristics such as difficulty in recalling words, grammar and syntactic errors, repetitive speech, and incoherent narratives, particularly when describing images. Thus, we used ChatGPT to assess the participants' image descriptions, querying for evaluation feedback, which we then leveraged for AD classification. Specifically, we first employed the Whisper ASR model [26] to transcribe text from the participants' speech signals. Subsequently, the transcribed text was input into ChatGPT along with a prompt to inquire about the fluency of the image descriptions, soliciting opinions from the LLM. Furthermore, the speech signals, transcribed text, and opinions were processed through a feature extractor composed of Wav2Vec 2.0 and BERT models, followed by a classification step that involved transformer blocks and linear layers to finally discern the possibility of AD in the participant.

The main contributions of this study are as follows. (i) An AD detection method is proposed using ChatGPT. (ii) The performance is compared by combining speech signals, transcribed text, and ChatGPT opinions using various approaches. (iii) Performance improvements are demonstrated when using both text and opinions as inputs, as confirmed through an ablation study.

The remainder of this paper is organized as follows. The dataset and related studies are presented in Sections 2 and 3, respectively. The proposed AD detection method is described in Section 4. Section 5 presents the performance evaluation of the proposed approach. The concluding remarks are presented in Section 6.

2 | DATASET

The ADReSSo Challenge provided benchmark datasets for three tasks: AD classification, MMSE score regression, and disease prognosis. The dataset for AD classification includes audio recordings of the "Cookie Theft" picture description task obtained from the Boston Diagnostic Aphasia Examination [14]. This task is commonly used to identify language disorders. Audio recordings were made for both cognitively healthy individuals and patients with AD. Participants were instructed to describe the "Cookie Theft" picture according to the guidelines of the Boston Diagnostic Aphasia Examination.

The dataset comprised 237 audio files. To address potential bias, the files were divided into training and test sets at a 70/30 ratio with careful consideration of sex and age distributions. The training set included 166 instances; 87 were diagnosed as AD patients (probable AD) and 79 as healthy elderly controls (control). The remaining 71 instances formed the test set, which consisted of 35 AD patients and 36 cognitively normal individuals.

3 | RELATED WORK

Over the past decade, research on dementia detection technologies that utilize speech as a biomarker has intensified [8]. Among these studies, the 2020 ADReSS Challenge [11] used both speech signals and transcribed text, while the 2021 ADReSSo Challenge [12] focused on speech signals, and the 2023 ADReSS-M Challenge [13] considered multilingual speech. In this context, the ADReSSo challenges, which are of interest in this study, are primarily aimed at detecting dementia using acoustic features extracted from audio [15–17] or word-embedding representations [9, 17, 18, 22] obtained from transcribed text acquired through ASR.

In [15], dementia was detected solely using acoustic features extracted from audio signals. They utilized conventional acoustic features such as fundamental frequency, jitter, and shimmer in conjunction with acoustic embeddings obtained using pretrained Wav2Vec 2.0. These two representations were concatenated and fed into a support vector machine (SVM) classifier, achieving an accuracy of 67.6% in their experiments. In [16], acoustic embeddings from various pretrained models, such as Trill [27], Allosaurus [28], and Wav2Vec 2.0 [23], were used for AD classification employing deep learning approaches. An experiment utilizing Wav2Vec 2.0 yielded an accuracy of 78.9%.

Text-based features exhibit superior performance compared with acoustic-based features [15, 16]. Specifically, in [22], the use of text-based embedding vectors with pretrained BERT achieved the highest accuracy reported to date at 84.5%, whereas the accuracy using acoustic features remained at 74.6%. Furthermore, in [18], specialized preprocessing for silent intervals within text information was introduced. Using this approach, text embedding for classification achieved an accuracy of 84.5%. These results suggest that text-based features outperform audio-based ones.

Recent research suggests that the integration of linguistic features with complementary attributes can significantly enhance the accuracy of AD prediction. In [19], a diverse range of acoustic features, including x-vectors, prosody, and emotional embeddings, were combined with word embeddings, resulting in an accuracy of 80.3%. In [20], an accuracy rate of 84.5% was achieved by incorporating x-vectors along with speaker information and word embeddings. Similarly, an earlier study [22] demonstrated the highest performance by fusing confidence scores from ASR with linguistic features, achieving an accuracy of 84.5%.

Furthermore, the artificial intelligence chatbot ChatGPT [25] has gained considerable popularity. ChatGPT is an LLM trained through reinforcement learning with human feedback and can generate realistic and accurate responses to user queries. However, to the best of our knowledge, no study has utilized the opinions of ChatGPT to aid in the early diagnosis of AD. Thus, we aim to contribute to AD detection by leveraging the opinions generated by ChatGPT.

4 | PROPOSED METHOD

In this section, we explain the method of performing AD recognition (ADR) from speech signals using ChatGPT responses. Figure 1 shows the entire process of the proposed ADR method, which includes the preprocessing, feature extraction, and classification.

4.1 | Data preparation

In this section, we describe data preparation, which involves obtaining transcribed text and evaluation statements from input audio. The ADReSSo Challenge provided audio files without transcribed text. These audio files were available to each participant with alternating recordings of the participants' and examiners' speeches. Organizers provided speaker labels and timestamps for each segment. However, the timestamps provided by organizers were inaccurate, and using both participant and examiner speech signals may aid in AD classification [16]. Therefore, we treated each audio file as a single sample without differentiating between participant and examiner speech signals.

4.1.1 | Speech2Text

We used the Whisper ASR system [26] to transcribe automatically text from audio files. The Whisper ASR system was trained for 680 000 h on multilingual speech data collected from the Web and demonstrated human-level recognition accuracy [26] for English speech. One transcript was obtained from each audio file. In this study, we refer to the automatically transcribed text as "text."

4.1.2 | Text2Opinion

We then utilized ChatGPT to analyze how fluent the participants described the images. ChatGPT generates appropriate responses when given a prompt related to a question. Therefore, an appropriate prompt must be defined. To define the prompt, we began with the statement that the task involved the analysis of descriptions of the "Cookie Theft" picture from the Boston Diagnostic Aphasia Examination. Subsequently, we structured the

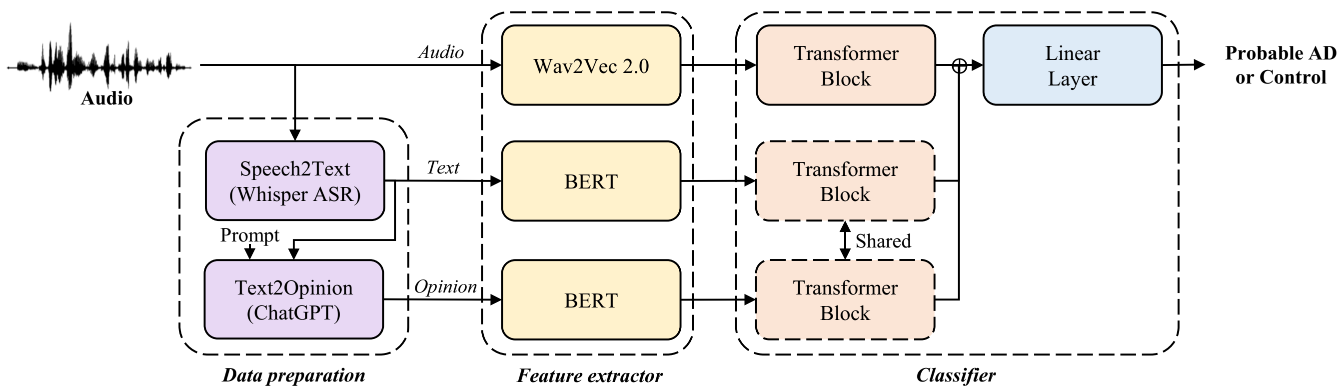


FIGURE 1 Overall process of proposed Alzheimer's disease recognition (ADR) method.

prompt by asking how fluent the participants described the given picture and requested an evaluation score between 1 and 10. To ensure accurate responses, sample prompts with the lowest and highest scores were included. These sample prompts were based on the MMSE scores provided by the ADReSSo Challenge [12], where the MMSE score represents the cognitive assessment score obtained using pen-and-paper tests. Finally, we designed a prompt using the desired text for evaluation, which was the automatically transcribed text of the audio, in which each participant described the picture. We refer to the response obtained from ChatGPT as "opinion." This opinion included the analysis results of fluency and clarity, including whether the participants mentioned the key elements of the picture, repetition, and unclear phrases in the context. Detailed examples of the prompts are provided in Section 5. Finally, we prepared the input data for ADR, which consisted of audio, text, and opinions from every participant.

4.2 | Feature extraction

We followed the approach used in [15–18], where high-dimensional embedding vectors were extracted using Wav2Vec 2.0 [23] and BERT [24]. First, we employed Wav2Vec 2.0 to extract the acoustic embedding vectors for each audio file. This framework was designed for self-supervised learning of raw audio representations. Previous studies [15, 16] have utilized this model to extract speech-based embedding vectors. We utilized the pretrained Transformers Python library Wav2Vec 2.0 base model [23] pretrained for 960 h on the Librispeech corpus at a sampling rate of 16 kHz. Like in [16], we segmented the audio files into 20 s segments for feature extraction because processing large audio files in the model is impractical. Consequently, the audio files for each participant yielded acoustic embeddings generated every 20 ms.

For the text and opinions, we employed the BERT model to extract language embedding vectors. Various studies on the ADReSSo Challenge [9, 12, 15–22] have reported impressive performance using the BERT model. Thus, the BERT model was adopted in this study. Specifically, we utilized the pretrained BERT tokenizer provided by the Transformers Python library to tokenize words and then employed the BERT model [24] to extract language embeddings for each token. Longer texts were divided into segments comprising up to 512 tokens for feature extraction. Most texts and opinions comprised fewer than 512 tokens, adhering to the defined token limit. Therefore, we did not apply a separate sliding-window technique to longer texts. Consequently, both text and opinions yielded language embeddings for each token.

Overall, we employed the BERT model to extract 768-dimensional embedding vectors per token from text and opinion data. Additionally, we used the Wav2Vec 2.0 model to extract 768-dimensional embedding vectors every 20 ms from audio data. The obtained embedding vectors were $\mathbf{x}_{\text{text}}$, $\mathbf{x}_{\text{opinion}}$, and $\mathbf{x}_{\text{audio}}$, and they were used as inputs for the classifier.

4.3 | Classification

The classifier utilizes embeddings extracted from diverse modalities of audio, text, and opinions to classify them into probable AD and control classes. This task is accomplished using a classifier composed of transformer blocks [10] and linear layers.

Input data $\mathbf{x}_{\text{text}}$, $\mathbf{x}_{\text{opinion}}$, and $\mathbf{x}_{\text{audio}}$ were processed using the same procedure within the classifier. For convenience, we denote this as $\mathbf{x}_{\text{mode}}$ and explain the process using the following equation. The embedding vectors for each modality $\mathbf{x}_{\text{mode}}$ are transformed into normalized data $\mathbf{x}_{\text{norm}}$, as shown in (1). Subsequently, normalized data x_{norm} are passed through transformer blocks specific to each modality, resulting in representative vectors suitable for classification. Equation (2) describes the self-attention mechanism emphasizing the correlations within the input data to obtain $\mathbf{x}_{\text{attn}}$. In (3), a feed-forward network is employed to further abstract the representation, yielding $\mathbf{x}_{\text{ffn}}$. LayerNorm$(\cdot)$ represents layer normalization, and Dropout$(\cdot)$ represents dropout.

$$\mathbf{x}_{\text{norm}} = \text{LayerNorm}(\mathbf{x}_{\text{mode}}), \tag{1}$$

$$\mathbf{x}_{\text{attn}} = \text{LayerNorm}(\mathbf{x}_{\text{norm}} + \text{Dropout}(\text{MHA}(\mathbf{x}_{\text{norm}}))), \tag{2}$$

$$\mathbf{x}_{\text{ffn}} = \text{LayerNorm}(\mathbf{x}_{\text{norm}} + \text{Dropout}(\text{FF}(\mathbf{x}_{\text{attn}}))). \tag{3}$$

The transformer blocks operate separately for each modality, as shown in Figure 1. However, because the text and opinion embeddings are based on BERT, they belong to the same modality. Therefore, we designed transformer blocks to share parameters across these embeddings.

The output of the variable-length transformer block is processed through average pooling AvgPool$(\cdot)$, as shown in (4). During this process, the embedding from each modality is transformed into a fixed-length vector denoted as $\overline{\mathbf{x}}_{\text{mode}}$. This vector has fixed dimensions $[768 \times 1]$ regardless of the input data sequence length.

$$\overline{\mathbf{x}}_{\mathrm{mode}} = \mathrm{AvgPool}(\mathbf{x}_{\mathrm{ffn}}). \tag{4}$$

The representative vectors of each modality, $\overline{\mathbf{x}}_{\mathrm{text}}$, $\overline{\mathbf{x}}_{\mathrm{opinion}}$, and $\overline{\mathbf{x}}_{\mathrm{audio}}$ generated using (1)–(4) are combined into a single vector $\mathbf{x}_{\mathrm{merged}}$ of dimensions $[2304 \times 1]$ through concatenation $\mathrm{Concat}(\cdot)$ as follows:

$$\mathbf{x}_{\mathrm{merged}} = \mathrm{Concat}\left(\overline{\mathbf{x}}_{\mathrm{text}}, \overline{\mathbf{x}}_{\mathrm{opinion}}, \overline{\mathbf{x}}_{\mathrm{audio}}\right). \tag{5}$$

Finally, classes *probable AD* and *control* are distinguished using a linear layer $\mathrm{Linear}(\cdot)$ as follows:

$$y = \mathrm{Linear}\left(\mathbf{x}_{\mathrm{merged}}\right). \tag{6}$$

The classifier training involved the utilization of cross-entropy loss $L = -\log p\left(y, \mathbf{x}_{\mathrm{text}}, \mathbf{x}_{\mathrm{opinion}} | \mathbf{x}_{\mathrm{audio}}\right)$. Consequently, the proposed model determined whether the given data belong to a probable AD or control case as the final diagnosis based on the provided text, opinions, and audio embedding vectors.

5 | EXPERIMENTS

5.1 | Experimental settings

All experiments were conducted using the PyTorch framework [29], and the following parameters were used. In data preparation, the large version of Whisper [26] was utilized as the speech recognizer, and ChatGPT version 4.0 was used [30]. In feature extraction, BERT and Wav2Vec employed a base model [31]. In classification, the attention size of the transformer was set to 768 with one attention head, and the feedforward model used 768 dimensions. The following settings were used: dropout rate of 0.1, learning rate of 0.01, and batch size of 8. A consistent random seed of zero was maintained across all the experiments, and training was performed for 50 epochs.

The experiments were conducted using fivefold cross-validation (CV). In the fivefold CV, the training data were divided into five segments, and each segment was sequentially employed as the validation data to form five folds. Within each fold, the model was trained and validated for 50 epochs, and the model with the highest accuracy was selected. Finally, the predictions of the five models selected from each fold were aggregated by voting to derive the final prediction. The evaluation was performed based on the accuracy, which measured the ratio of correctly classified samples to the total number of samples.

We employed various measures of the classification performance using the Scikit-learn Python API [32], namely, precision, recall, F-score, specificity, and accuracy. We considered the accuracy as the main measure for performance comparison, as in the ADReSSo Challenge [12].

5.2 | Opinion analysis

In this section, we present examples of opinions obtained from ChatGPT. Examples provide the input prompts and corresponding responses generated by ChatGPT for both control and probable AD cases.

Figure 2 shows the prompts and responses of ChatGPT for a control sample. This figure first provides input prompts for evaluating descriptions of the "Cookie Theft" image. These prompts included two responses (1 and 10 points). The 1-point response reflects a

Example prompt for a Control sample

This is an example describing the "Cookie Theft" photo from the Boston Diagnostic Aphasia Exam. Could you please evaluate the fluency of this description and provide a rating on a scale of 1 to 10? Below, you can find examples corresponding to ratings of 1 and 10.

Example of a 1-point rating: "What's going on in the picture? In here? Mm-hmm. This way? Mm-hmm. I really don't know because I haven't gotten things turned up. Have a look at the picture, see if you can tell me what's going on. This one. I think it's right here. What is it? Oh my God. That isn't... This isn't kind of... Whatever. What is it that's happening here? She's cleaning. And what are they doing? Looking to see what they have so they can get it ready for bed. I don't know. I hear it's down. It's falling down. I don't know. I hear it's down. It's falling down."

Example of a 10-point rating: "I'd like you to tell me everything that you see going on in the picture. And that's the picture. Action type things. There is water coming out of a faucet into a basin which is overflowing onto the floor. Mother is drying dishes and she's walking in the water that is flowing on the floor. Two children are stealing cookies. One of whom is up on a stool that is about to fall. He's removing cookies from an upper shelf in the closet. The wind appears to be blowing because the curtains appear to be puffed. Those are the only action things that I can see."

Here is an example: "Just the action. Well, the <u>girl</u> is reaching for a cookie that the <u>boy</u> is trying to get for her. While he's... Am I going too fast? While he's falling off the ladder. And the <u>mother</u> is washing dishes, drying a plate. While the <u>sink</u> is spilling over with water that is overflowing from the spigot that is still running. And <u>water</u> is gushing out. And the <u>girl</u> is signaling to be quiet. Not to lift another note that the boy is getting a cookie for her. And I did say that <u>she's</u> drying the dish with the towel. And the <u>water</u> splashing on the floor."

Generated Response

"This description is quite thorough, with many of the key elements in the picture noted. It accurately mentions the children trying to get a cookie, the boy's unstable footing, the mother washing dishes, the sink overflowing with water, and even minor details like the girl signaling for quiet and the mother drying a dish with a towel. There are no repeated statements or major inaccuracies, and it also adds some speculation on the characters' motivations and actions, which could indicate a deeper level of understanding. One point of confusion might be the phrase Not to lift another note that the boy is getting a cookie for her', which seems out of context or possibly miscommunicated. If 'note' was meant to refer to something specific in the image, it's unclear what that could be. Overall, the description is detailed and organized, albeit with one confusing phrase. If this is an error or miscommunication, correcting this could improve the score. Considering these aspects, this example could be rated **around 8 or 9 out of 10**. It provides a comprehensive, coherent, and generally accurate account of the events in the picture, with one potentially confusing phrase being the main detraction."

* Each example in the prompt is a transcription obtained from an ASR and may contain transcription errors.

FIGURE 2 Prompt and answer for control sample.

participant's brief description with a lack of clear understanding of the image. The 10-point response thoroughly describes various details within the image. Speech recognition results for the control sample were then input into the prompts.

The responses generated from ChatGPT for prompts evaluated the participants' descriptions, highlighting their strengths and potential confusion or inaccuracies. The generated responses were typically rated at eight or nine points, reflecting a high level of comprehension and relevance to the image in the participants' descriptions. These responses provided insights into the model's evaluation process and the quality and accuracy of the descriptions provided.

Each example in the prompt is a transcription obtained from ASR and may contain transcription errors.

Each example in the prompt is a transcription obtained from ASR and may contain transcription errors.

Figure 3 shows the ChatGPT prompts and answers for a probable AD sample. This figure shows the application of ChatGPT for evaluating the descriptions of participants with probable AD. The input prompt began with the same description as the control sample and includes 1- and 10-point sample responses, as shown in Figure 2.

An example response from this sample illustrates a participant with probable AD attempting to describe an image. The generated responses noted the disjointed nature of the description and the absence of a coherent narrative of the image. Repetitions and phrases that lack contextual meaning are evident in the participant's description ("Girl. Girl. Girl. Girl.," "Garage. Garage. Garage. Garage."). The speaker appeared to struggle to identify and express the elements of the scene, which is consistent with the characteristics of aphasia often observed in individuals with AD. Important details such as the mother drying dishes, overflowing sink, boy on the stool, and cookies were not mentioned. The rating of the generated response for this description was 2 out of 10, emphasizing the challenges faced by individuals affected by AD in expressing themselves coherently and accurately.

The box plot in Figure 4 shows a comparison of the descriptive scores obtained from ChatGPT for the control and probable AD classes. This graph was obtained using a training set from the ADReSSo corpus. The *x* axis shows the classes control and probable AD, and the *y* axis shows the descriptive scores assigned by ChatGPT. The control class demonstrated a higher average descriptive score (6.9) than the probable AD class (4.4). These results highlight the potential utility of descriptive scores obtained from ChatGPT for classification.

5.3 | Unimodal results

In this section, we compare the AD classification performance of each modality, namely, audio, text, and opinion. Table 1 lists the AD classification accuracy of the unimodal methods. Baselines A and B represent

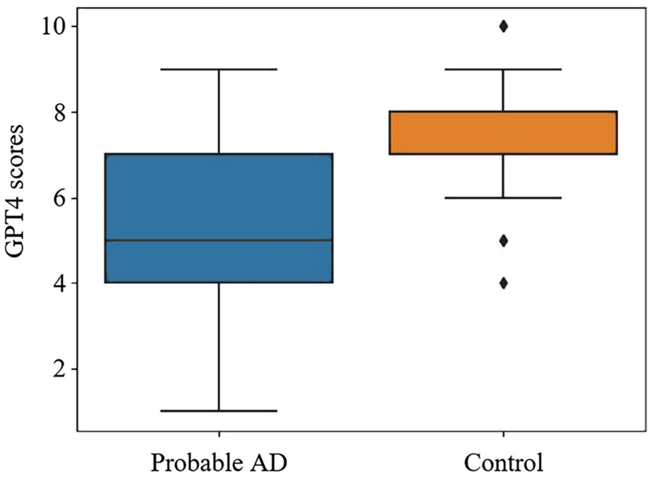


FIGURE 4 Comparison of ChatGPT scores for control and probable Alzheimer's disease (AD) cases.

Example prompt for probable AD sample

This is an example describing the "Cookie Theft" photo from the Boston Diagnostic Aphasia Exam. Could you please evaluate the fluency of this description and provide a rating on a scale of 1 to 10? Below, you can find examples corresponding to ratings of 1 and 10.

Example of a 1-point rating: "What's going on in the picture? …

Example of a 10-point rating: "I'd like you to tell me everything that …

Here is an example: "Have a look at that picture and tell me everything that you see going on in that picture, all the action, what's happening there? Flying through the courtyard. I'm sorry? Flying through the courtyard. Going through the courtyard. They're all falling. Girl. They're fishing. Boy. Riding God. Gun void. Pardon? Gun void. Right. Girl. Pardon? Girl. Girl. Girl. Girl. Pardon? What's happening here? Boy. Girl. Boy. Good. Anything else? Pardon? Garage. I'm sorry? Garage. Garage. Garage. Garage. Okay. Good. Okay. Okay. Good."

Generated Response

"This description is quite disjointed and doesn't provide a clear narrative of what is happening in the picture. It contains a lot of repetitions (Girl. Girl. Girl. Girl.', 'Garage. Garage. Garage. Garage.') and phrases that don't make sense in context ('Flying through the courtyard.', 'Riding God. Gun void.'). The speaker seems to struggle with identifying what's happening in the image or expressing it in words, which is typical in people with aphasia. Many key elements of the scene are not mentioned, such as the mother drying dishes, the sink overflowing, the boy on the stool, and the cookies. Based on these observations, this description is quite low on fluency and clarity. I would rate this description as a **2 out of 10**."

* Each example in the prompt is a transcription obtained from an ASR and may contain transcription errors.

FIGURE 3 Prompt and answer for probable Alzheimer's disease (AD) sample.

TABLE 1 Alzheimer's disease (AD) classification results of unimodal methods.

Method	Modality	Pr %	Rc %	Fs %	Sp %	Ac %
Baseline A [12]	Text	–	–	–	–	77.46
Baseline B [12]	Audio	–	–	–	–	64.79
Baseline C [15]	Audio	63.64	80.00	70.89	–	67.61
Proposed (unimodal)	Text	*83.10*	*83.10*	*83.10*	*83.33*	*83.10*
	Audio	70.60	70.42	70.39	66.67	69.01
	Opinion	71.72	70.42	69.88	83.33	70.42

Note: Baselines A and B show results measured using LOSO-CV.
Abbreviations: Ac, accuracy; Fs, F-score; LOSO-CV, leave-one-subject-out cross-validation; Pr, precision; Rc, recall; Sp, specificity.

TABLE 2 Alzheimer's disease (AD) classification results of multimodal methods.

Method	Modality	Pr %	Rc %	Fs %	Sp %	Ac %
Baseline D [12]	Text–audio	–	–	–	–	78.90
Baseline E [22]	Text + ASR conf.	–	–	–	–	84.51
Proposed (multimodal)	Text–audio	85.87	84.51	84.33	94.44	84.51
	Text–opinion	86.94	85.92	85.80	94.44	85.92
	Audio–opinion	81.13	80.28	80.12	88.89	80.28
	Text–opinion–audio	85.87	84.51	84.33	94.44	84.51
	Text–opinion–audio + shared	*88.06*	*87.32*	*87.25*	*94.44*	*87.32*

Note: Baseline D shows the results obtained through late fusion using LOSO-CV. Baseline E provides the state-of-the-art performance achieved in the past ADReSSo Challenge [12].
Abbreviations: Ac, accuracy; ASR conf., ASR confidence score; Fs, F-score; LOSO-CV, leave-one-subject-out cross-validation; Pr, precision; Rc, recall; Sp, specificity.

the accuracies of the baseline models provided by the ADReSSo Challenge [12]. Baseline A achieved an accuracy of 77.5% by extracting linguistic features from the speech recognition results using EVAL and FREQ commands in the Computerized Language Analysis (CLAN) program [33], followed by SVM classification. Baseline B achieved an accuracy of 64.8% by extracting acoustic features based on the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPs) [34] and SVM classification. Both experiments were conducted using a leave-one-subject-out (LOSO) CV [12] instead of a fivefold CV. In contrast, baseline C [12] achieved an accuracy of 67.6% by using the Mel-Frequency Cepstral Coefficient (MFCC)- and Wav2Vec-2.0-based acoustic features for SVM classification. These baseline experimental results indicated that linguistic features were more beneficial than acoustic features for AD classification.

In the proposed method, we verified the classification performance with unimodal data by omitting the concatenation step in the classification stage shown in Figure 1. Our experiment utilized text features and achieved high accuracy (83.1%). The discrepancy in performance between our proposed method and baseline A, which also used text features, could be attributed to our utilization of the pretrained BERT model to extract specialized features and improved speech recognition performance achieved through Whisper ASR. The experiments utilizing audio features resulted in an accuracy of 69.0%, which closely matched those of baselines B and C. Furthermore, the opinions extracted from ChatGPT showed similar accuracy outcomes as the audio features at 70.4%.

5.4 | Multimodal results

We compared the AD classification accuracies of the multimodal methods. To this end, we input different combinations of text, opinion, and audio information into the classifiers, as listed in Table 2. In Table 2, "shared" indicates the sharing of the transformer block parameters of Figure 1.

In Table 2, baseline D [12] represents the multimodal classification performance provided by the organizers. They achieved a 78.9% LOSO-CV accuracy by late-fusing acoustic and language information. Baseline E [22] achieved a state-of-the-art performance of 84.5% in the ADReSSo Challenge by utilizing a pretrained BERT model and an ASR confidence score based on speech recognition.

TABLE 3 Alzheimer's disease (AD) classification results according to opinion types.

Opinion type	Pr %	Rc %	Fs %	Sp %	Ac %
GPT 4.0	*71.72*	*70.42*	*69.88*	83.33	*70.42*
GPT 3.5	66.47	61.97	58.86	*88.89*	61.97
Shuffle	50.72	50.70	50.70	50.00	50.70

Abbreviations: Ac, accuracy; Fs, F-score; Pr, precision; Rc, recall; Sp, specificity.

TABLE 4 Alzheimer's disease (AD) classification results according to ASR performance.

ASR	Pr %	Rc %	Fs %	Sp %	Ac %
Whisper [26]	*83.10*	*83.10*	*83.10*	*83.33*	*83.10*
Wav2Vec 2.0 [23]	81.69	81.94	81.69	77.78	81.67

Abbreviations: Ac, accuracy; ASR, automatic speech recognition; Fs, F-score; Pr, precision; Rc, recall; Sp, specificity.

We evaluated the AD classification performance using various combinations of text–audio, text–opinion, and audio–opinion modalities. When combining text and audio, we achieved an improved accuracy of 84.5%, which was comparable to that of baseline D and may be attributed to the influence of Whisper ASR. Baseline E [22] employed Wav2Vec 2.0 ASR. The performance gaps owing to the different ASR systems are discussed in Section 5.6. Furthermore, the combination of text and opinion yielded higher accuracy (85.9%) than the combination of text and audio. This indicates that the opinion modality contributed valuable information to AD classification. The combination of audio and opinion modalities showed an accuracy of 80.3%, surpassing the accuracy of the audio-based approach but falling short of the accuracy achieved by the text-based approach. This finding underscores the effectiveness of text modalities for AD classification.

When utilizing the text, opinion, and audio modalities together, we achieved an accuracy of 84.5%, which was lower than that of the text- and opinion-based approaches. This could be attributed to an increase in the dimensionality of the features handled by the classifier from 1536 to 2304 dimensions.

However, based on an experiment in which the transformer blocks of the text and opinion modalities were shared, a maximum accuracy of 87.3% was achieved. This outcome could be attributed to the shared pretrained BERT model for both modalities and the increased training data for the transformer blocks. Conversely, when sharing the transformer block parameters of the audio modality, a decline in AD classification performance was observed. Overall, we developed an AD classifier that leverages opinion information from ChatGPT and achieved an accuracy of 87.3%.

5.5 | Ablation study

We investigated the effects of the opinion quality on AD classification. To this end, we generated three types of opinions, as shown in Table 3. (i) Opinions from GPT version 4.0. These opinions were the same as those used in the previous unimodal and multimodal experiments. (ii) Opinions from GPT version 3.5. (iii) Shuffled opinions, which were obtained by randomly shuffling the opinions obtained from GPT version 4.0.

Experiments were conducted in a unimodal environment. Opinions obtained from GPT version 4.0 exhibited a high accuracy of 70.4%. In contrast, those obtained from GPT version 3.5 yielded a lower accuracy of 62.0%. This suggests a decline in the ability of ChatGPT to evaluate image descriptions in previous versions. Finally, shuffled opinions yielded an extremely low accuracy of 50.7%. These results can be attributed to the decreased correlation between shuffled opinions and correct labels. Thus, we confirmed that the quality of opinions generated by ChatGPT influenced AD classification.

5.6 | Effects of ASR performance on AD classification

We also investigated the effects of the ASR system performance on AD classification, as shown in Table 4. We compared the recently proposed Whisper ASR [26] with the Wav2Vec 2.0 ASR [23] used in [22]. It has been reported that there is a significant difference [26] in English speech recognition performance between Wav2-Vec 2.0 and Whisper ASR.

Experiments were conducted using hypothesized texts generated from each speech recognition system for AD classification in a unimodal environment. The results demonstrate that the hypothesized texts generated using Whisper ASR achieved an AD classification accuracy of 83.1%. By contrast, the hypothesis texts obtained from Wav2Vec 2.0 ASR exhibited a relatively low accuracy of 81.7%. This underscores the significant impact of the ASR performance on speech-based AD classification.

6 | CONCLUSION

We propose a method to automatically predict AD from speech data using ChatGPT. We demonstrate that utilizing ChatGPT to analyze participants' picture description abilities is effective for AD classification. We compared AD classification performance by combining speech signals, transcribed text, and ChatGPT opinions in various ways, thus demonstrating that incorporating opinion information from different modalities substantially contributes to performance enhancement. The method that uses audio, text, and opinions achieved the highest performance of 87.3%.

Furthermore, we confirmed the effect of ChatGPT-generated opinions on AD classification through an ablation study. We deduced that the choice of generative model could notably affect the final classification accuracy. This indicates the potential for using more advanced generative models to offer better outcomes in AD classification. In future work, we will explore methods using generative models to derive AD classification results solely from natural language rather than from just picture descriptions.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

ORCID

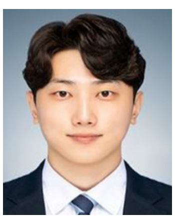
Jeong-Uk Bang https://orcid.org/0000-0002-0439-6802
Byung-Ok Kang https://orcid.org/0009-0001-8217-720X

REFERENCES

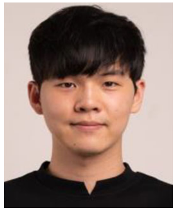
1. R. Li and Y. Liu, *Physical activity and prevention of Alzheimer's disease*, J. Sport Health Sci. **5** (2016), 381–382.
2. M. F. Folstein, S. E. Folstein, and P. R. McHugh, *"Mini-mental state": a practical method for grading the cognitive state of patients for the clinician*, J. Psychiatr. Res. **12** (1975), 189–198.
3. Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow, *The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment*, J. Am. Geriatr. Soc. **53** (2005), 695–699.
4. S. Chen, D. Stromer, H. A. Alabdalrahim, S. Schwab, M. Weih, and A. Maier, *Automatic dementia screening and scoring by applying deep learning on clock-drawing tests*, Sci. Rep. **10** (2020), DOI 10.1038/s41598-020-74710-9
5. I. Vigo, L. Coelho, and S. Reis, *Speech-and language-based classification of Alzheimer's disease: a systematic review*, Bioengineering **9** (2022), DOI 10.3390/bioengineering9010027
6. S. Dong and H.-B. Jeon, *Feature analysis and evaluation for estimation of mild cognitive impairment from the spontaneous speech of Korean*, (Proc. International Congress on Acoustics, Gyeongju, Republic of Korea), 2022.
7. E. Hussain, M. Hasan, S. Z. Hassan, T. H. Azmi, M. A. Rahman, and M. Z. Parvez, *Deep learning based binary classification for Alzheimer's disease detection using brain MRI images*, (15th IEEE Conference on Industrial Electronics and Applications—ICIEA, Kristiansand, Norway), 2020, pp. 1115–1120.
8. S. de la Fuente Garcia, C. W. Ritchie, and S. Luz, *Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review*, J. Alzheimers Dis. **78** (2020), 1547–1574.
9. J. Chen, J. Ye, F. Tang, and J. Zhou, *Automatic detection of Alzheimer's disease using spontaneous speech only*, (Proceedings of Interspeech, Brno, Czechia), 2021, pp. 3830–3834.
10. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, *Attention is all you need*, (31st International Conference on Neural Information Processing Systems—NIPS, Long Beach, CA, USA), 2017, pp. 6000–6010.
11. S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, *Alzheimer's dementia recognition through spontaneous speech: the ADReSS Challenge*, (Proceedings of Interspeech, Shanghai, China), 2020, pp. 2172–2176.
12. S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, *Detecting cognitive decline using speech only: the ADReSSo Challenge*, (Proceedings of Interspeech, Brno, Czechia), 2021, pp. 3780–3784.
13. S. Luz, F. Haider, D. Fromm, I. Lazarou, I. Kompatsiaris, and B. MacWhinney, *Multilingual Alzheimer's dementia recognition through spontaneous speech: a signal processing grand challenge*, arXiv preprint (2023), DOI 10.48550/arXiv.2301.05562
14. H. Goodglass and E. Kaplan, *The Boston diagnostic aphasia examination*, Lea & Febinger, Philadelphia, 1983.
15. Balagopalan Aparna and J. Novikova, *Comparing acoustic-based approaches for Alzheimer's disease detection*, arXiv preprint (2021), DOI 10.48550/arXiv.2106.01555
16. L. Gauder, L. Pepino, L. Ferrer, and P. Riera, *Alzheimer disease recognition using speech-based embeddings from pre-trained models*, (Proceedings of Interspeech, Brno, Czechia), 2021, pp. 3795–3799.
17. Y. Zhu, A. Obyat, X. Liang, J. A. Batsis, and R. M. Roth, *Wavbert: exploiting semantic and non-semantic speech using Wav2vec and BERT for dementia detection*,(Proceedings of Interspeech, Brno, Czechia), 2021, pp. 3790–3794.
18. Z. S. Syed, M. S. Syed, M. Lech, and E. Pirogova, *Tackling the ADRESSO Challenge 2021: the MUET-RMIT system for Alzheimer's dementia recognition from spontaneous speech*, (Proceedings of Interspeech, Brno, Czechia), 2021, 3815–3819.
19. N. Wang, Y. Cao, S. Hao, Z. Shao, and K. P. Subbalakshmi, *Modular multi-modal attention network for Alzheimer's disease detection using patient audio and language data*, (Proceedings of Interspeech, Brno, Czechia), 2021, pp. 3835–3839.
20. R. Pappagari, J. Cho, S. Joshi, L. Moro-Velázquez, P. Zelasko, J. Villalba, and N. Dehak, *Automatic detection and assessment of Alzheimer's disease using speech and language technologies in low-resource scenarios*, (Proceedings of Interspeech, Brno, Czechia), 2021, pp. 3825–3829.
21. P. A. Pérez-Toro, S. P. Bayerl, T. Arias-Vergara, J. C. Vásquez-Correa, P. Klumpp, M. Schuster, E. Nöth, J. R. Orozco-Arroyave, and K. Riedhammer, *Influence of the interviewer on the automatic assessment of Alzheimer's disease in the context of the ADReSSo Challenge*, (Proceedings of Interspeech, Brno, Czechia), 2021, pp. 3785–3789.
22. Y. Pan, B. Mirheidari, J. M. Harris, J. C. Thompson, M. Jones, J. S. Snowden, D. Blackburn, and H. Christensen, *Using the outputs of different automatic speech recognition paradigms for acoustic-and BERT-based Alzheimer's dementia detection through spontaneous speech*, (Proceedings of Interspeech, Bron, Czechia), 2021, pp. 3810–3814.

23. A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, *Wav2vec 2.0: a framework for self-supervised learning of speech representations*, In *Advances in neural information processing systems*, 2020, 12449–12460.
24. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, *BERT: pre-training of deep bidirectional transformers for language understanding*, (Proceedings of NAACL-HLT, Minneapolis, MN, USA), 2019, pp. 4171–4186.
25. OpenAI, *ChatGPT*, 2023, Available from: https://chat.openai.com/chat, [last accessed July 2023].
26. A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *Robust speech recognition via large-scale weak supervision*, (International Conference on Machine Learning, Honolulu, HI, USA), 2023, pp. 28492–28518.
27. J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. D. Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv, *Towards learning a universal non-semantic representation of speech*, arXiv preprint (2020), 10.48550/arXiv.2002.12764
28. X. Li, S. Dalmia, J. Li, M. Lee, P. Littell, J. Yao, A. Anastasopoulos, D. R. Mortensen, G. Neubig, A. W. Black, and F. Metze, *Universal phone recognition with a multilingual allophone system*, (IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain), 2020, pp. 8249–8253.
29. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and A. Desmaison, *PyTorch: an imperative style, high-performance deep learning library*, (Advances in Neural Information Processing Systems, Vancouver, Canada, 2019.
30. OpenAI, *GPT-4 technical report*, arXiv preprint (2023), arXiv: 2303.08774.
31. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Davison, *Transformers: state-of-the-art natural language processing*, 2020, pp. 38–45.
32. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, *Scikit-learn: machine learning in Python*, J. Mach. Learn. Res. **12** (2011), 2825–2830.
33. B. MacWhinney, *Tools for analyzing talk part 2: the CLAN program*, Talkbank, 2017.
34. F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, *The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing*, IEEE Trans. Affect. Comput. **7** (2015), 190–202.

AUTHOR BIOGRAPHIES



Jeong-Uk Bang received the BS degree in electronics engineering, MS degree in control and instrumentation engineering, and PhD degree in control and robot engineering from Chungbuk National University, Cheongju, Republic of Korea, in 2013, 2015, and 2020, respectively. From 2020 to 2022, he was a postdoctoral researcher at the Electronics and Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea. He is currently a senior researcher at the Superintelligence Creative Research Laboratory, ETRI, Daejeon, Republic of Korea. His research interests include speech recognition, speech translation, backchannel prediction, and Alzheimer's disease investigation.



Seung-Hoon Han received the BS degree in German language and literature from Chungnam National University, Daejeon, Republic of Korea, in 2021. He is currently pursuing the MS degree with the Department of Brain and Cognitive Engineering, Korea University, Seoul, Republic of Korea and is a student researcher at the Superintelligence Creative Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea. His research interests include speech recognition, cognitive sciences, and healthcare.

Byung-Ok Kang received the BS and MS degrees in electronics and electrical engineering from POSTECH, Pohang, Republic of Korea, in 1997 and 1999, respectively. He also received the PhD degree in control and robot engineering from Chungbuk National University, Cheongju, Republic of Korea, in 2018. He worked at Samsung Electronics from 1999 to 2001 and joined the Integrated Intelligence Research Section at the Electronics and Telecommunications Research Institute in 2001. His research interests include automatic speech recognition, speech processing, unsupervised and semisupervised learning, and tutoring and medical artificial intelligence.

How to cite this article: J.-U. Bang, S.-H. Han, and B.-O. Kang, *Alzheimer's disease recognition from spontaneous speech using large language models*, ETRI Journal **46** (2024), 96–105. DOI 10.4218/etrij.2023-0356.