Master's Thesis of Cognitive Science

A performance comparison of the speech tasks: to discover the optimal task for automated speech-based Alzheimer's disease detection

발화과제 성능비교: 자동화된 발화기반 알츠하이머병 탐지를 위한 최적의 발화과제 탐색

August 2022

Graduate School of Humanities
Seoul National University
Interdisciplinary Program in Cognitive Science

Minju Bae

A performance comparison of the speech tasks: to discover the optimal task for automated speech-based Alzheimer's disease detection

Examiner Jun-Young Lee

Submitting a master's thesis of
Cognitive Science

August 2022

Graduate School of Humanities
Seoul National University
Interdisciplinary Program in Cognitive Science

Minju Bae

Confirming the master's thesis written by
Minju Bae
August 2022

Chair	Kyogu Lee	(Seal)
Vice Chair	Jun-Young Lee	(Seal)
Examiner	Seyul Kwak	(Seal)

Abstract

A performance comparison of the speech tasks: to discover the optimal task for automated speech-based Alzheimer's disease detection

Graduate School of Humanities

Seoul National University

Interdisciplinary Program in Cognitive Science

Minju Bae

Objective: Voice is one of the promising markers which facilitates the early screening of Alzheimer's disease (AD). Previous studies in automatic speech-based AD detection generally focused on the improvement of accuracy in AD classification by the refinement of algorithms, and rarely investigated the optimal speech task

which induces and captures the distinguishing acoustic features of AD voice. In the present study, we suggest several speech tasks which imposes cognitive load to participants and evaluate the potential of speech tasks as an automatic speech-based AD detection method.

Methods: The present study collected speech recordings from 79 AD patients and 79 healthy controls using three speech tasks: Interview, Repetition, and Recall. The interview task consisted of 5 questions about participants' daily life. The repetition task and recall task were carried out using two modified well-known fairy-tales. In the repetition task, participants were asked to listen and repeat the given story phrase-by-phrase following a researcher. In the recall task, participants were asked to recall the new-learned information of modified well-known stories as specific as possible. Speech recordings were segmented into single utterances. We built separate AD classification models and cognitive impairment prediction models with speech datasets from each speech task: Interview, Repetition, and Recall. Features to be used to build models were selected by analysis of variance ($p < 0.005$). In AD classification, Random Forest (RF), Support Vector Machine (SVM), Naive-Bayes (NB), and k-Nearest Neighbor (k-NN) were used and in cognitive impairment prediction, RF, SVM, and Ridge were used.

Results: In AD classification, the best performing model was the RF model trained on the speech dataset from the recall task which reported a CV accuracy of 72.9%. The models trained on the speech dataset from the recall task outperformed speech datasets from other speech tasks regardless of the used classifiers. In cognitive impairment prediction, the best performing model was the SVM model trained on

the speech dataset from the recall task which achieved a CV RMSE of 5.34 and a CV MAE of 4.38. Likewise, the speech dataset collected from the recall task achieved the best accuracy regardless of the used regressors.

Conclusions: The present study confirms that the performance of AD classification and cognitive impairment prediction can be influenced by the speech task used to collect speech data. Among three speech tasks, Interview, Repetition, and Recall, used in the present study, the recall task seems to have superiority over other speech tasks in AD classification and cognitive impairment prediction. The present study suggests the cognitive load imposed by the recall task might affect the speech production mechanism and induces the distinguishing acoustic feature of AD patients. For future works, it is necessary to focus on exploring the optimal task which reflects the characteristic of AD voice abundantly for automatic speech-based AD detection.

Keyword: Alzheimer's disease, Mini Mental State Examination, Speech Acoustics, Supervised Machine Learning

Student Number: 2020-17753

Table of Contents

List of Tables

List of Figures

List of Supplementary Materials

Chapter 1. Introduction

In the aging society we face today, dementia is one of the most serious threats to older adults' health and quality of life. In 2019, about 55.2 million people worldwide live with dementia and the number is increasing rapidly (GBD, 2021). It is expected that in 2030, about 78 million people will live with dementia and in 2050, about 139 million will do (GBD, 2021). The social cost caused by dementia is tremendous. The estimated global cost of dementia approached about 1.3 trillion USD in 2019 (WHO, 2021). In order to relieve the burden we bear now, the early diagnosis and intervention of dementia are crucial.

According to the amyloid cascade hypothesis (Karran, Mercken, & Strooper, 2011), the molecular change such as the deposition of the amyloid-β peptide and tau protein in Alzheimer's disease (AD) occur before observable clinical symptoms such as cognitive. Therefore, when patients recognize cognitive decline and visit the clinic, the pathology of dementia must have progressed considerably already. This is why when it comes to dementia, the early diagnosis is the key to successful treatment. However, existing diagnosis methods probably might not be the best solutions for early diagnosis due to their invasive, expensive, and time-consuming aspects.

Therefore, new diagnosis methods based on voice markers are getting attention in order to overcome the limitations of previous diagnosis methods. Also, the fact that language impairment is a prominent symptom in patients with AD supports the innovation. For instance, AD patients tend to struggle with word finding difficulties in their early phase (Slegers, Filiou, Montembeault, & Brambati, 2018). As the

pathology progresses, patients have difficulty in understanding the conversation, and even repeat a certain sound, word, or sentence (Klimova, Maresova, Valis, Hort, & Kuca, 2015). Severe AD patients show impairment in comprehension, reading, and writing (Ferris & Farlow, 2013). In spontaneous speech, these language impairments manifest in the acoustic features of voice. Thus, lots of studies attempted to apply the characteristic acoustic features of voice to diagnose AD (Balagopalan & Novikova, 2021; Luz, Haider, de la Fuente, Fromm, & MacWhinney, 2021; Yuan et al., 2020).

Most previous studies in this area used speech data from well-established datasets such as Pitt Corpus (Becker, Boiler, Lopez, Saxton, & McGonigle, 1994; MacWhinney, 2019), BEA Hungarian dataset (Gósy, 2013), Gothenburg MCI database (Wallin et al., 2016), the Carolina Conversation Collection (CCC) (Pope & Davis, 2011). Various kinds of speech tasks were employed for speech data collection. While the most frequently used task was the Cookie Theft Picture Task (Goodglass & Kaplan, 1983), the semantic verbal fluency task (Benton, Hamsher, & Sivan, 1976), or an unstructured interview (Pope & Davis, 2011) were also commonly used. Other researchers used the language-related subtests of batteries to measure cognitive abilities or intelligence, such as Wechsler Adult Intelligence Scale (WAIS-III) (Wechsler, 1997a), Wechsler Memory Scale (WMS-III) (Wechsler, 1997b). (Table 1.)

However, to the best of our knowledge, previous studies focused on the improvement of accuracy in AD classification and cognitive impairment prediction by refinement of algorithms, there was little effort to explore the optimal task which captures the distinctive speech characteristics of AD patients effectively and propose a novel task for automatic speech-based AD detection. Most previous studies used existing tasks which were developed for cognitive evaluation or unstructured

interview and conversation to collect speech data from AD patients. Also, there was a relatively small number of studies which attempt to compare the performance and potential of speech tasks as new diagnosis methods for automatic AD detection. Even though datasets contain speech samples from several tasks, most previous studies did not distinguish them from each other and used the entire speech data to automatically detect AD. However, it is important to understand the intrinsic characteristics of speech tasks and find the optimal speech task which captures the distinguishing alterations in AD because different speech tasks demand different functions and abilities. Also, required functions can be influenced by AD pathology and show significant differences between AD patients and healthy older adults, or less influenced by AD pathology and rather preserved nevertheless and make it difficult to observe the alterations in speech characteristics. For instance, the picture description task facilitates the assessment of the lexico-semantic level (March, Wales, & Pattison, 2006), while it requires limited syntactic ability, and responses are mainly restricted to simple constructions (Garrard & Forsyth, 2010). Otherwise, the interview is used to elicit spontaneous speech and it is useful in analyzing discourse-pragmatic ability, syntactic and semantic processing (Lai, 2014) (Ripich, Carpenter, & Ziol, 2000) (Sajjadi, Patterson, Tomek, & Nestor, 2012).

The previous researchers observed that increased cognitive load imposed by cognitive tasks affects speech production and causes the change in acoustic features of voice. Physiological change of cepstral peak prominence and low-to-high spectral energy ratio was observed in the voice of healthy young adults when a cognitive load was imposed by the Stroop task (MacPherson, Abur, & Stepp, 2017). Increased

Table 1. Available speech datasets for automatic speech-based AD detection

Database	Participants	Speech Task	Language
Pitt Corpus	• Healthy Adults • Probable AD	• The Cookie Theft Picture Description	English
BEA Hungarian Dataset	• Healthy Adults	• Spontaneous speech	Hungarian
Gothenburgh MCI Database	• MCI patients	• The Cookie Theft Picture Description • Reading task	Swedish
Carolina Conversation Collection	• Older patients with chronic conditions	• Conversation	English

cognitive load also affects the stability and timing of speech, yet the effect was greater in older adults than in younger adults (MacPherson, 2019). A previous study presented the possibility that the decay in speech motor performance in older adults might be affected by an age-related cognitive decline (MacPherson et al., 2017) (Tremblay et al., 2018). In keeping with previous studies, we can speculate that speech task which imposes more cognitive load would have an impact on the speech production in AD patients and induce the subtle change of the acoustic features of voice.

The modified well-known fairy-tale recall task is one of the story recall tasks which provides participants a revised old famous fairy-tale and asks them to recall the story as accurately as possible. This task requires an ability to suppress the retrieval of the well-known fairy tale story from long-term memory and recall the new-learned information (Attali, De Anna, Dubois, & Barba, 2009; De Anna et al., 2008). In the previous study, researchers provided healthy controls and AD patients three kinds of stories: a new story, a well-known fairy-tale, and a modified well-known fairy-tale. Healthy controls did not show significant differences in recall accuracy between stimuli, while AD patients were interfered by over-learned information from the original story and showed more errors in recall (De Anna et al., 2008). If cognitive load imposed by a modified well-known fairy-tale story recall task in the form of response inhibition captures specific voice patterns of AD patients, it might be a new effective diagnosis method for automatic speech-based AD detection.

In this study, we build classification models and prediction models to screen AD patients from healthy controls and predict the severity of cognitive impairment using speech data collected from three different speech tasks. We suggest a novel speech

task which amplifies distinguishing voice patterns of AD patients by applying a modified well-known fairy-tale recall task. Also, we compare the model performances trained on each dataset collected from different speech tasks to explore the optimal speech task for automatic speech-based AD detection.

Chapter 2. Methods

2.1. Participants

The data included a collection of speech recordings from 79 people with AD and 79 healthy controls. AD patients were recruited from SMG-SNU Boramae Medical Center for Dementia. The clinical diagnosis of AD was based on the National Institute of Aging and the Alzheimer's Association (NIA-AA) criteria (Jack Jr et al., 2018). Subjects suspected or diagnosed with dementia types other than AD were not included in the analysis, including vascular dementia, Lewy body dementia, frontotemporal lobe dementia, and vascular dementia. Healthy controls were recruited from Dongjak center for dementia in Seoul, Korea. Healthy controls were screened based on Mini-Mental Statement Examination (MMSE) (MMSE score ≥ 27) (Cockrell & Folstein, 2002). All participants were over 65 years of age and we matched sex, education years between groups (Table 2).

Table 2. Demographics of participants in the dataset. AD, Alzheimer's disease; HC, Healthy controls; MMSE, Mini-Mental Statement Examination score; p-value, independent t-test or Chi-squared test were used as appropriate.

	AD (n=79)	HC (n=79)	p-value
Sex, female (%)	46 (58.2)	49 (62.0)	0.626
Mean Age, y (SD)	80.38 (5.54)	74.34 (2.37)	0.000
Mean Education, y (SD)	7.57 (5.05)	8.34 (4.48)	0.311
MMSE score (SD)	17.39 (4.00)	28.71 (1.16)	0.000

2.2. Speech task

In this study, three different types of speech tasks were administered: Interview, Repetition, and Recall. In the interview task, participants were interviewed about their daily life and personal information. The interview was composed of 5 questions which ask about participants' age, education years, yesterday's dinner menu, recently watched TV program, and how they felt recently (Supplementary Material 1).

In the repetition and recall tasks, the modified well-known fairy-tale recall task was utilized. We selected two renowned fairy-tales, "Kongji and Patji" and "The tale of Sim Cheong" (Supplementary Material 2) (김하나 & 성지은, 2014). We revised the relationship between characters, the personality of characters, and the ending of the story compared to the original story. Each fairy-tale was composed of 8 sentences. Two fairy-tales were read aloud to the participants in a counterbalanced order.

In the repetition task, participants listened and repeated the fairy-tale phrase-by-phrase following the researcher. If the participants could not repeat the phrase correctly, the researcher repeated the phrase up to one time. After finishing the repetition, participants were asked to recall the story immediately as specific as possible. If the response was not enough or missed key information, the researcher encouraged the participant to describe more in detail up to two times.

The participant's speech was recorded using a smartphone as m4a sound files and converted to wav sound files after recording.

2.3. Data Preparation

The acquired speech samples were segmented into single utterances manually. The definition of utterance remains unclear in previous studies. In this study, we defined utterance as a continuous piece of speech that is preceded by silence or a researcher's speech and followed by a change of speaker or completion of a sentence. Twenty researchers who were trained about audio segmentation guidelines manually tagged the time stamp of corresponding utterances using open-source software, 'Voice studio 2.0' (Figure 1). Four other researchers checked the quality of time stamp tagging. 30 samples were excluded because of extreme noise or sound distortion.

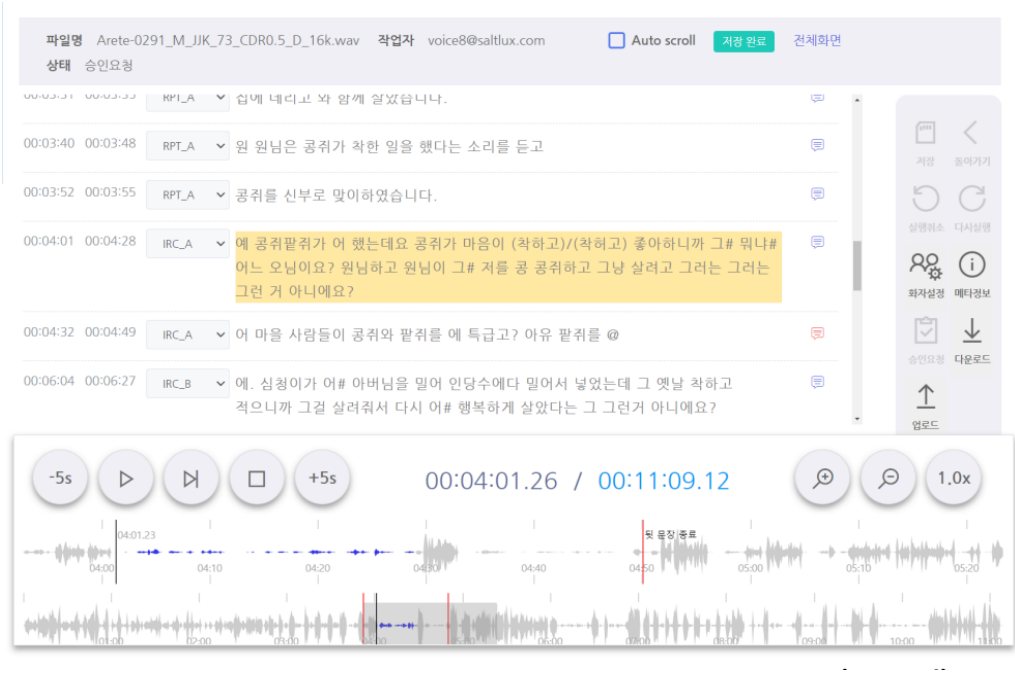

Figure 1. The software for audio segmentation, "Voice studio 2.0".

In the present study, we extracted the eGeMAPS feature set (Eyben et al., 2015) using the Python library openSMILE toolkit (Eyben, Wöllmer, & Schuller, 2010). The eGeMAPs is the minimalistic set of voice parameters that was originally developed to recognize the affective state of a speaker based on the information that the voice conveys but recently, the eGeMAPS is widely used in various areas of speech analysis, especially in automatic speech-based AD detection (Haider, de la Fuente, Albert, & Luz, 2020; Haider, De La Fuente, & Luz, 2019; Pappagari et al., 2021; Valsaraj, Madala, Garg, & Baths, 2021). The reason why the eGeMAPS is commonly used in many studies is that it consists of a standardized, limited set of features, which were chosen based on their theoretical relevance and potential to analyze important aspects of speech (Xue, Cucchiarini, van Hout, & Strik, 2019). The eGeMAPS is divided into four parameter groups:

Frequency-related, Energy/Amplitude-related, Spectral, Temporal. Details of each parameter group are presented in Table 3.

Table 3. Four parameter groups of eGeMAPS

Parameter groups	Features
Frequency-related	• Pitch • Jitter • Formant 1,2,3
Energy/Amplitude-related	• Shimmer • Loudness • Harmonics-to-noise-ratio (HNR)
Spectral	• MFCC • Spectral Flux • Alpha Ratio • Hammarberg Index • Spectral Slope
Temporal	• the mean length and the standard deviation of voiced regions • the rate of loudness peaks • the number of continuous voiced regions per second

2.4. AD Classification

After data preparation, features to be used to build AD classification models were selected based on the analysis of variance (ANOVA) to reduce the high dimensionality of feature space. Among 88 features which were originally

extracted from the eGeMAPS, statistically significant features ($p < 0.005$) were selected to build classification models.

Unlike previous studies which compared the performances of different classifiers in AD detection, this study compares the performances of datasets that were collected using different speech tasks to explore the optimal speech task which induces distinctive acoustic patterns of AD voice for automatic speech-based AD detection. Yet we used four classifiers to guarantee that the outperformance of a certain dataset is not attributed to the competence or intrinsic difference of a classifier. We used four different classifiers which are commonly used to classify AD patients from healthy older adults using speech: Random Forest (RF), Support Vector Machine (SVM), Naive Bayes (NB), k-Nearest Neighbor (k-NN) (de la Fuente Garcia, Ritchie, & Luz, 2020). All classifiers used the default hyperparameter values.

Datasets were split into 80%-20% for training and testing sets. Because each recording from participants was segmented into utterances, the dataset consists of several speech segments from single participants. When we split datasets into training and test sets, we ensured that there were no speaker overlaps while preserving a similar class distribution in each split set. All models were trained on the training set and 10-fold cross-validation (CV) was performed on training set. To evaluate and compare the performances of different speech tasks, testing and validation accuracy, AUC, precision, recall, and F1-Score were measured.

Grouped feature importance analysis was performed using an leave-one-out analysis (Sankaranarayanan et al., 2021) (Hartanto, Sami, de Ridder, & Nijveen, 2022). Each feature was categorized into five groups; Frequency-related, Amplitude-related, Temporal, Spectral, and MFCC based on the feature description

(Eyben et al., 2015). Mel frequency Cepstrum Coefficients (MFCC) is one of the representative spectral features that is based on the human peripheral auditory system (Tiwari, 2010). To prevent certain feature groups having too many number of features and accidentaly getting high feature importance score, we separated MFCC related features from spectral features and balanced the size of the feature group. Each feature group was iteratively excluded from the dataset, and a model was trained using the reduced dataset. Feature importance was computed as the AUC difference between the full model with all features and reduced model.

2.5. Cognitive Impairment Prediction

We built cognitive impairment prediction models in the analogous procedure with AD classification models. Features were selected to build cognitive impairment prediction models based on the ANOVA ($p < 0.005$). Three different regressors were employed to predict the MMSE score: RF, SVM, and Ridge. All regressors used the default hyperparameter values. All models were trained on the training set and 10-fold CV was performed on training data. We calculated the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE) to evaluate the performance of cognitive impairment prediction models (Chai & Draxler, 2014a). Feature importance was evaluated as the difference in the CV-RMSE between the full model and the reduced model.

Chapter 3. Results

3.1. Dataset Description

For the classification of AD patients and healthy controls and prediction of cognitive impairment, we collected speech recordings from participants using three different speech tasks (i.e. Interview, Repetition, Recall). Table 4 summarizes the amount of speech data available after audio segmentation.

Table 4. Duration of available data for the AD classification and cognitive impairment prediction in hours (# segments). AD, Alzheimer's disease; HC, Healthy Controls.

Speech Task	AD	HC	Total
Interview	1:21 (1236)	1:25 (1343)	2:46 (2579)
Repetition	1:45 (1489)	1:29 (1358)	3:14 (2847)
Recall	1:19 (775)	1:27 (528)	2:46 (1303)

3.2. AD Classification

The classification results of each speech task are shown in Table 5. We measured AUC and two kinds of accuracy to evaluate the distinguishability of the classification models. The area under the Receiver Operating Characteristics (ROC) curve, AUC is the generally used standard method to assess the performance of classification models (Huang & Ling, 2005). Also, we measured

testing accuracy and CV accuracy. Testing accuracy is the accuracy of the model at predicting the test data and CV accuracy is the average of the ten estimates using 10-fold CV which is measured to assess the generalizability of a model to unseen data (Tabe-Bordbar, Emad, Zhao, & Sinha, 2018). Hence in this study, we will focus on the AUC and CV accuracy for comparisons between the classification models.

The results in Table 5 show that the RF model trained on the recall dataset reported the best performance with 72.9% CV accuracy and 0.82 AUC. The SVM, NB, and k-NN models that were trained on the recall dataset also provided better results than the models trained on the interview task or repetition datasets achieving 69.5%, 71.6%, 71.6% CV accuracy, and 0.75, 0.70, 0.77 AUC respectively. We can see that the recall dataset yielded the best classification performances among three speech tasks regardless of the classifiers.

Table 5. A comparison of the performance of AD classification models. The best accuracy is given in bold. ACCU, accuracy; CV, Cross-validation; RF, Random Forest; SVM, Support Vector Machine; NB, Naive Bayes; k-NN, k-Nearest Neighbors.

ACCU	Task	RF	SVM	NB	k-NN
CV-ACCU	Interview	69.2%	68.2%	67.7%	66.5%
	Repetition	65.6%	66.9%	65.6%	66.9%
	Recall	**72.9%**	**69.5%**	**71.6%**	**71.6%**
Testing ACCU	Interview	67.2%	67.5%	63.4%	64.5%
	Repetition	61.3%	60.8%	56.4%	60.5%
	Recall	**71.4%**	**69.0%**	**65.5%**	**69.4%**
AUC	Interview	0.71	0.72	0.60	0.67
	Repetition	0.68	0.67	0.60	0.64
	Recall	**0.82**	**0.75**	**0.70**	**0.77**

The second-best performing speech task was the interview task. The RF, SVM, and NB models trained on the interview dataset reports better CV accuracy compared to the models trained on the repetition dataset with 68.2%, 69.2%, and 67.7%, respectively. In the case of AUC, the RF, SVM, k-NN models trained on the interview dataset achieved better AUC than the repetition dataset with 0.71, 0.72, 0.67 AUC.

For further insight, the confusion matrices of the best results of each speech task (i.e. Interview, Repetition, Recall) are also shown in Figure 2. The best results were decided based on the CV accuracy and AUC score and the following models were reported; the SVM model trained on the interview dataset, the RF model trained on the repetition dataset, and the RF model trained on the recall dataset. Precision refers to the probability that participants who were classified as AD patients indeed are AD patients, and negative predicted value refers to the probability that participants who were classified as healthy controls indeed are healthy controls. Sensitivity refers to the ability to detect AD patients, specificity refers to the ability to distinguish healthy controls from AD patients (Trevethan, 2017). F1-Score is a measure that evaluates the performance of models combining both precision and sensitivity while the accuracy measures how many observations were correctly classified. The results in Figure 2. show that the RF model trained on the recall dataset provides the best performance with a 77.5% F1-Score. Specifically, the RF model with recall dataset reported the highest precision and sensitivity compared to the interview or repetition dataset with 72.0% and 84.0%, respectively.

Feature importance analysis demonstrates how much each feature contributed to distinguishing AD patients and helps interpret the model. Since we had a large

feature set of 88 features, we categorized features into five groups reflecting their properties and evaluated the importance of the feature group. Figure 3 shows that classification models heavily rely on spectral features across all tasks. Spectral features include spectral flux, alpha Ratio, and Hammarberg Index which characterize voice timbre, a particular attribute of voice which differentiate voice from other (Cleveland, 1977). In the model with a recall dataset, the next-highest scoring feature group was frequency features which determine the pitch of voice, whereas in other models with interview and repetition datasets, amplitude features got the second highest feature importance score.

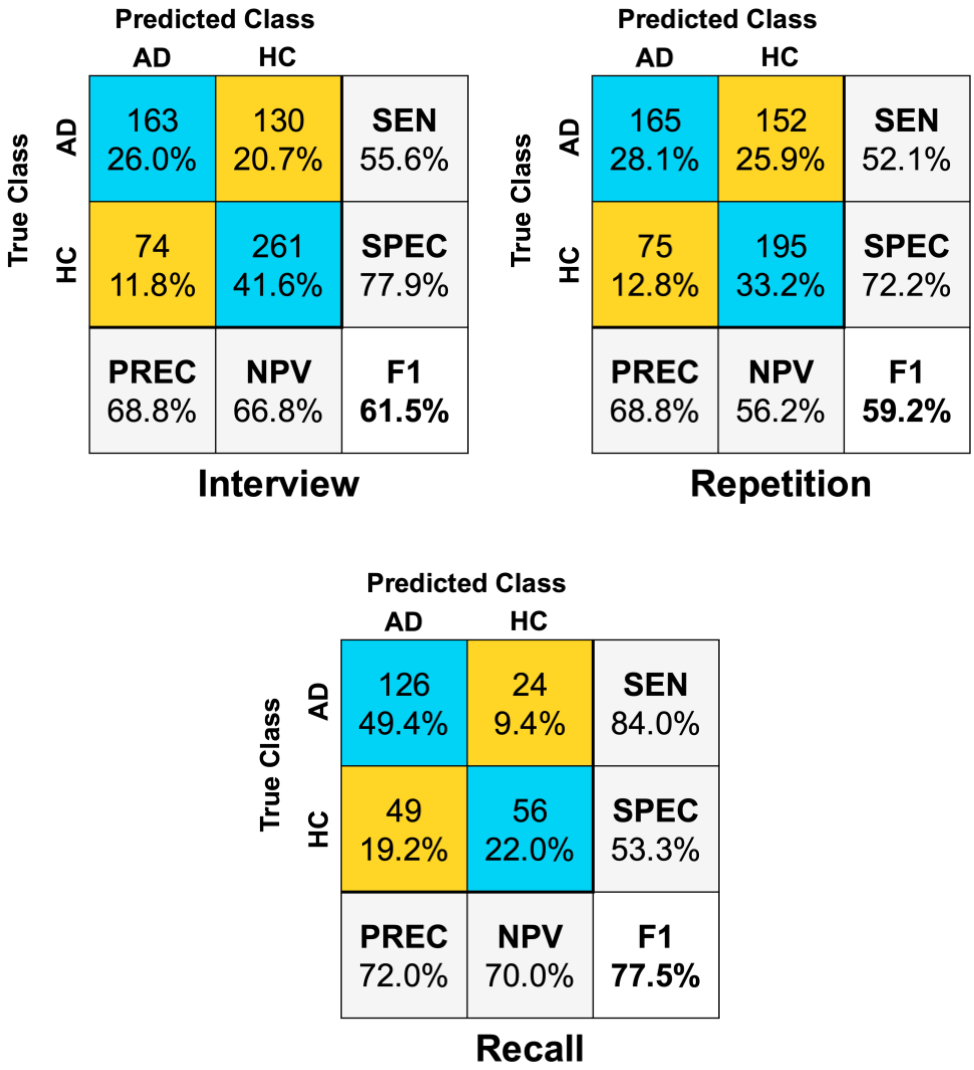

Figure 2. The confusion matrices of the best results of each speech task. AD; Alzheimer's Disease, HC, Healthy Controls; PREC, Precision; NPV, Negative predictive value; SEN, Sensitivity; SPEC, Specificity; F1, F1-score.

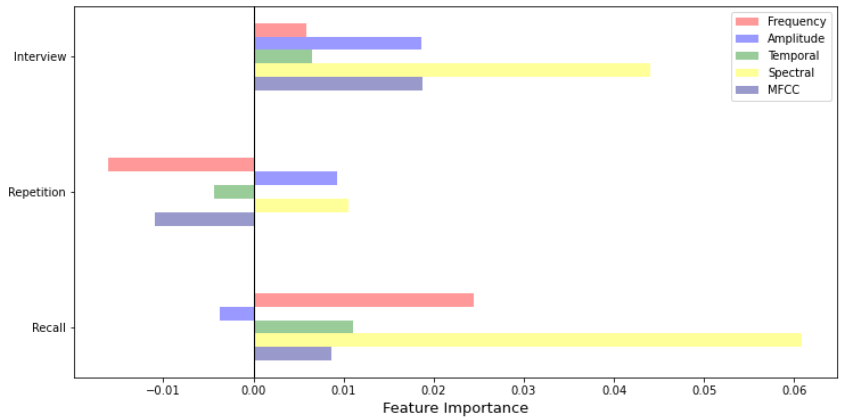

Figure 3. Feature importance by task and feature group in AD classification model. Feature importance score is defined by the drop in the area under the receiver operator characteristic curve when each feature group is removed from the analysis.

3.3. Cognitive Impairment Prediction

Table 6 provides cognitive impairment prediction results. To evaluate the performances of models and compare them we computed Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) on the testing set and validation set. The RMSE and MAE are metrics which generally used to measure how the estimates are accurate. As these metrics are smaller, the predictability of the model is better. Although it is controversial which metric is a better indicator of model performance, since both are generally used in model evaluation, in this study we computed both metrics (Chai & Draxler, 2014b).

The results in Table 6 show that the SVM model trained on the recall dataset reported the best performance with 5.34 CV RMSE and 4.38 CV MAE. As with AD classification results, the cognitive impairment prediction models trained on

the recall dataset reported the best result compared to the models trained on other speech task datasets irrespective of regressors.

Table 6. A comparison of the performance of cognitive impairment prediction models. The best result is given in bold. ACCU, accuracy; CV, Cross-validation; RMSE, Root Mean Square Error; MAE, Mean Absolute Error, RF, Random Forest; SVM, Support Vector Machine.

ACCU	Task	RF	SVM	Ridge
CV-RMSE	Interview	6.01	6.02	5.94
	Repetition	6.25	6.08	6.79
	Recall	**5.62**	**5.34**	**5.94**
CV-MAE	Interview	5.06	4.91	5.01
	Repetition	5.23	4.91	5.38
	Recall	**4.70**	**4.38**	**4.86**
Testing RMSE	Interview	5.30	5.45	5.94
	Repetition	5.21	5.41	5.59
	Recall	**4.96**	**5.13**	**5.46**
Testing MAE	Interview	4.58	4.68	5.28
	Repetition	4.60	4.69	4.86
	Recall	**4.17**	**4.31**	**4.69**

For further insight, we drew a boxplot of the absolute errors between the observed MMSE score and the predicted MMSE score of the best prediction models trained on the datasets from each speech task in Figure 3. The best prediction models were decided based on CV RMSE and CV MAE and the following models were reported: the SVM model trained on the interview dataset, the SVM model trained on the repetition dataset, and the SVM model trained on the recall dataset. Although the prediction models were built based on the MMSE

scores without knowing the diagnosis of participants, we computed the absolute errors separating the groups of participants to understand the predictability of each model deeply. From Figure 3, we can observe that the prediction model trained on the recall dataset did not report a statistically significant difference in absolute errors between groups (p-value = 0.83). On the other hand, prediction models that trained on the interview dataset and repetition dataset showed significant differences in absolute error between groups. We can see that both models trained on the interview dataset and repetition dataset had difficulty in the prediction of MMSE score of AD patients and produced larger errors in AD patients than in healthy controls.

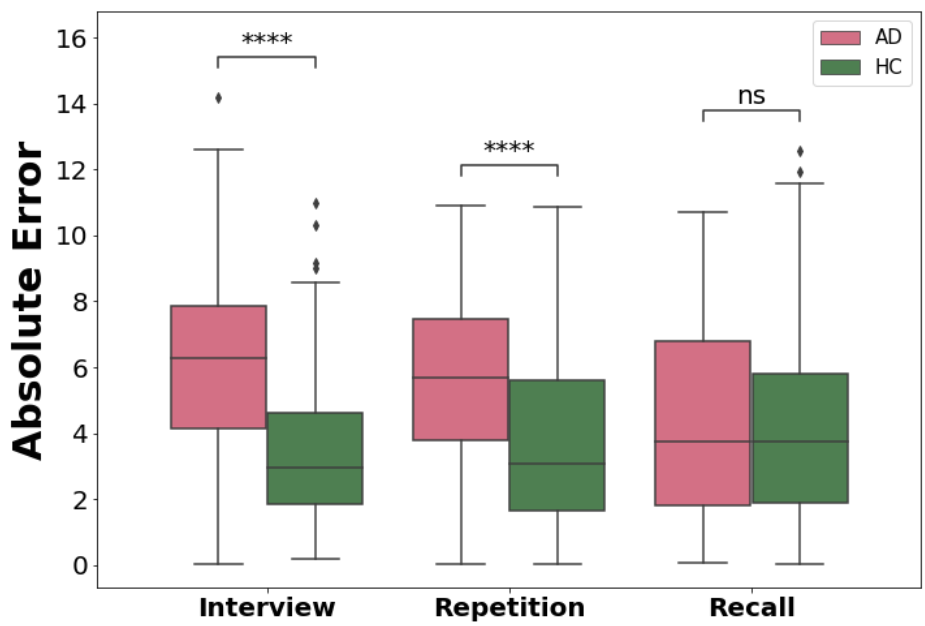

Figure 4. The absolute errors of the best prediction models trained on each speech task dataset. HC, Healthy Controls; AD, Alzheimer's Disease; ****, p-value ≤ 1.00e-04; ns, non-significant; Independent-test was used for comparison between groups.

Figure 5 shows the feature importance analysis in the cognitive impairment prediction model. We can observe that the pattern of feature importance score differs from the one in the AD classification model. Instead of spectral features, frequency-related features were the highest scoring features in two models with repetition dataset and recall dataset in the prediction of cognitive impairment. Also, while in the model with repetition dataset, all feature groups except spectral features relatively got high feature importance score, only one feature group was noticeable in the model with interview and recall dataset.

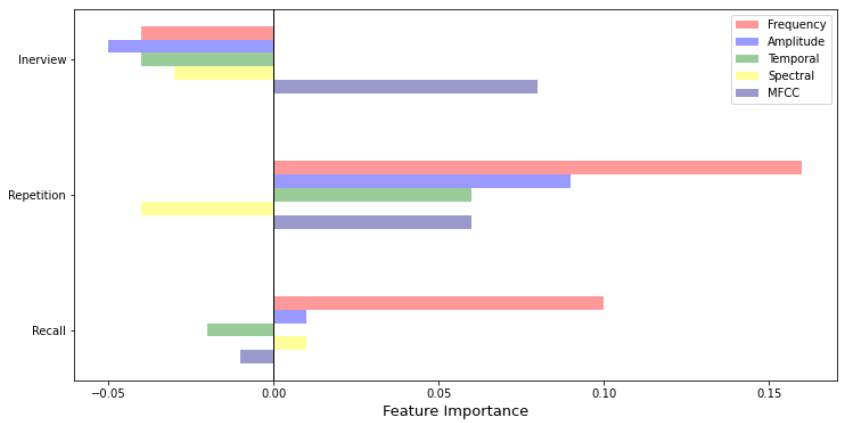

Figure 5. Feature importance by task and feature group in cognitive impairment prediction model. Feature importance score is defined by the increase in the Root Mean Square Error (RMSE) when each feature group is removed from the analysis.

Chapter 4. Discussion and Conclusion

4.1. Discussion

In this study, we compared the AD classification and cognitive impairment prediction performance of three different speech tasks to discover the optimal task that induces and captures distinguishing vocal features of AD patients and evaluate the potential as an automatic speech-based AD detection method. We conducted three different speech tasks (i.e. Interview, Repetition, and Recall) and built classification models which screen AD patients from healthy controls and cognitive impairment prediction models which estimate MMSE score with the collected speech data from each speech task.

We found that the recall task showed better classification performance compared to other speech tasks (i.e. Interview, Repetition) with 72.9% CV accuracy. The recall task also showed higher precision and sensitivity with 72.0% and 84.0%, respectively. These results highlight the optimality of recall task as an automatic speech-based AD detection method. Especially, it is noteworthy that the model trained on the recall dataset showed superior sensitivity (84.0%) relative to the interview and repetition task (52.6%, 52.1%, respectively). Sensitivity is considered important in the diagnosis method since low sensitivity indicates a high risk of missing an AD patient. As mentioned above, because early diagnosis and intervention of AD is very crucial in successful treatment, high sensitivity is

desirable for AD diagnosis methods. In this respect, we can say that the recall task has merit over other speech tasks.

In the cognitive impairment prediction, likewise, the prediction model trained on the recall dataset showed excellent predictability with 5.34 CV RMSE and 4.38 CV MAE. Also, the model trained on the recall dataset did not report extra difficulty in the prediction of the MMSE score of AD patients, whereas the models trained on the interview and repetition datasets showed low predictability in AD patients and produced larger errors predicting the MMSE score of AD patients than healthy controls. These results emphasized that the recall task not only has superiority in distinguishing AD patients from healthy controls, but also in predicting the severity of cognitive impairment by estimating the exact MMSE score.

Previous studies in automatic detection of AD using voice mainly focused on the improvement of model performance based on the refinement of the algorithm and rarely focused on the importance of the choice of the adequate speech method and data acquisition methods. As the result of the comparisons between three speech tasks in this study, we could find out the recall task induces the particular attribute of AD voice and leads to the improvement of model performance. The recall task we used in this study increases the cognitive load by demanding participants to inhibit the retrieval of over-learned information from long-term memory and retrieve the new-learned information. In many studies, the link between cognition and speech production has been addressed. The increased cognitive load might influence the speech motor mechanism and lower the speech stability which manifests into particular attributes of AD patients. These results are supported by previous findings about cognitive load and speech movement. It has

been reported that as the cognitive demands of the task increase, especially when inhibitory functioning is required, it impacts speech stability negatively (Dromey & Benson, 2003) (Dromey & Shim, 2008) (Whitfield, Holdosh, Kriegel, Sullivan, & Fullenkamp, 2021). When participants have reduced cognitive resources due to aging or neurodegenerative disease, they were more influenced by the increased cognitive load. Under the cognitive load condition, both older adults and young adults showed a change in the timing and characteristics of speech, but the effect was larger in older adults (MacPherson, 2019). Also, when the same cognitive load was imposed, whereas healthy older adults showed the same level of performance, AD patients showed a reduced level of performance and more performance errors (De Anna et al., 2008). In this study, since AD patients have reduced cognitive resources and they were vulnerable to the influences of the high cognitive load of the recall task, the distinguishing acoustic features of AD voice might be induced and maximize the difference in acoustic features of speech between AD patients and healthy older adults.

4.2. Limitations

The limitations of this study are as follows. First of all, we matched the sex ratio and mean education year between AD patients and healthy controls, but we could not match the age of participants. Aging entails the physiological alterations in the larynx, vocal cords, and articulation mechanisms (Mueller, 1997) (Etter et al., 2019) (Lindström, Öhlund Wistbacka, Lötvall, Rydell, & Lyberg Åhlander, 2022).

We are not certain that voice alterations due to age difference in the present study were significant since we only recruited participants over 65 years of age, yet it would be able to yield more reliable results if we match age between two groups in future works.

Secondly, the size of the datasets was relatively small in this study. For instance, the Pitt corpus, the most widely used speech dataset, contains 307 speech recordings from 194 AD patients and 242 speech recordings from 99 non-AD participants while we collected speech data from 79 AD patients and 79 healthy controls. Since large datasets contain more abundant information and assure the model performances to some degree, collecting more datasets would help models learn and capture the distinguishing voice characteristics of AD patients.

The last limitation of the study is that we extracted the acoustic features of voice using an existing conventional feature set. Even though it is true that the eGeMAPS is consist of standardized features with a reliable theoretical base and a generally used feature set in automatic speech-based AD detection, since the eGeMAPS was originally developed to recognize emotional states of a speaker, it might not be the best feature set to describe the AD voice. For instance, the eGeMAPS contains a limited set of temporal features. Among 88 acoustic features which the eGeMAPS contains, only six features are temporal features (Eyben et al., 2015). However, in the previous studies, it is commonly known that temporal features, like pause, play a significant role in automatic speech-based AD detection because AD influences the temporal features of speech (Hoffmann et al., 2010) (Szatloczki, Hoffmann, Vincze, Kalman, & Pakaski, 2015) (Pistono et al., 2016) (Pastoriza-Domínguez et al., 2022). Therefore, in future works, it would be

important to not only explore the optimal speech task but also to invent the optimal acoustic features that describe the AD voice accurately.

4.3. Conclusions

Taken together, the present study evaluated the competence of three speech tasks as an automatic speech-based AD detection method to explore the optimal task which captures the specific features of AD voice. While previous studies usually used speech data from existing speech datasets and rarely focused on which speech task should be used to collect speech datasets, the present study emphasizes the importance of careful speech task selection which can reflects the characteristics of language impairment of AD in automatic speech-based AD detection. This study confirms that AD classification and cognitive impairment prediction performance can be influenced by speech tasks and there are differences in potential as an automatic speech-based AD detection method between speech tasks. Among three speech tasks, interview, repetition, and recall tasks which used a modified well-known fairy-tale, the recall task reported the best performance in AD classification and cognitive impairment prediction. Our findings may suggest that the cognitive load imposed by the modified well-known fairy-tale recall task in the form of response inhibition affects speech production and causes the specific voice pattern of AD patients. We hope that the present study can be a contribution to development of a novel AD-specific speech task for the reliable and convenient automatic speech-based AD detection method which facilitates the early diagnosis of AD.

References

Attali, E., De Anna, F., Dubois, B., & Barba, G. D. (2009). Confabulation in Alzheimer's disease: poor encoding and retrieval of over-learned information. *Brain, 132*(1), 204-212.

Balagopalan, A., & Novikova, J. (2021). Comparing Acoustic-based Approaches for Alzheimer's Disease Detection. *arXiv preprint arXiv:2106.01555*.

Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., & McGonigle, K. L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of neurology, 51*(6), 585-594.

Benton, A., Hamsher, K. d., & Sivan, A. (1976). Multilingual Aphasia Examination. Iowa City. *University of Iowa*.

Chai, T., & Draxler, R. R. (2014a). Root mean square error (RMSE) or mean absolute error (MAE). *Geoscientific Model Development Discussions, 7*(1), 1525-1534.

Chai, T., & Draxler, R. R. (2014b). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific model development, 7*(3), 1247-1250.

Cleveland, T. F. (1977). Acoustic properties of voice timbre types and their influence on voice classification. *The Journal of the Acoustical Society of America, 61*(6), 1622-1629.

Cockrell, J. R., & Folstein, M. F. (2002). Mini-mental state examination. *Principles and practice of geriatric psychiatry*, 140-141.

De Anna, F., Attali, E., Freynet, L., Foubert, L., Laurent, A., Dubois, B., & Dalla

Barba, G. (2008). Intrusions in story recall: When over-learned information interferes with episodic memory recall. Evidence from Alzheimer's disease. *Cortex, 44*(3), 305-311.

de la Fuente Garcia, S., Ritchie, C. W., & Luz, S. (2020). Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review. *Journal of Alzheimer's Disease, 78*(4), 1547-1574.

Dromey, C., & Benson, A. (2003). Effects of concurrent motor, linguistic, or cognitive tasks on speech motor performance.

Dromey, C., & Shim, E. (2008). The effects of divided attention on speech motor, verbal fluency, and manual task performance.

Etter, N. M., Hapner, E. R., Barkmeier-Kraemer, J. M., Gartner-Schmidt, J. L., Dressler, E. V., & Stemple, J. C. (2019). Aging Voice Index (AVI): reliability and validity of a voice quality of life scale for older adults. *Journal of Voice, 33*(5), 807. e807-807. e812.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., . . . Narayanan, S. S. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing, 7*(2), 190-202.

Eyben, F., Wöllmer, M., & Schuller, B. (2010). *Opensmile: the munich versatile and fast open-source audio feature extractor.* Paper presented at the Proceedings of the 18th ACM international conference on Multimedia.

Ferris, S. H., & Farlow, M. (2013). Language impairment in Alzheimer's disease and benefits of acetylcholinesterase inhibitors. *Clinical interventions in aging, 8*, 1007.

Garrard, P., & Forsyth, R. (2010). Abnormal discourse in semantic dementia: A data-driven approach. *Neurocase, 16*(6), 520-528.

GBD. (2021). *Global Burden of Disease Study 2019 (GBD 2019)*. Retrieved from https://www.who.int/publications/i/item/9789240033245

Goodglass, H., & Kaplan, E. (1983). *Boston diagnostic aphasia examination booklet*: Lea & Febiger.

Gósy, M. (2013). BEA–A multifunctional Hungarian spoken language database. *Phonetician, 105*, 50-61.

Haider, F., de la Fuente, S., Albert, P., & Luz, S. (2020). Affective speech for Alzheimer's dementia recognition. *LREC: Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments (RaPID)*, 67-73.

Haider, F., De La Fuente, S., & Luz, S. (2019). An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing, 14*(2), 272-281.

Hartanto, M., Sami, A. A., de Ridder, D., & Nijveen, H. (2022). Prioritizing Candidate eQTL Causal Genes in Arabidopsis using Random Forests. *bioRxiv*.

Hoffmann, I., Nemeth, D., Dye, C. D., Pákáski, M., Irinyi, T., & Kálmán, J. (2010). Temporal parameters of spontaneous speech in Alzheimer's disease. *International journal of speech-language pathology, 12*(1), 29-34.

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering, 17*(3),

299-310.

Jack Jr, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., . . . Karlawish, J. (2018). NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia, 14*(4), 535-562.

Karran, E., Mercken, M., & Strooper, B. D. (2011). The amyloid cascade hypothesis for Alzheimer's disease: an appraisal for the development of therapeutics. *Nature Reviews Drug Discovery, 10*(9), 698-712. doi:10.1038/nrd3505

Klimova, B., Maresova, P., Valis, M., Hort, J., & Kuca, K. (2015). Alzheimer's disease and language impairments: social intervention and medical treatment. *Clinical interventions in aging, 10*, 1401.

Lai, Y.-h. (2014). Discourse features of Chinese-speaking seniors with and without Alzheimer's disease. *Language and Linguistics, 15*(3), 411-434.

Lindström, E., Öhlund Wistbacka, G., Lötvall, A., Rydell, R., & Lyberg Åhlander, V. (2022). How older adults relate to their own voices: a qualitative study of subjective experiences of the aging voice. *Logopedics Phoniatrics Vocology*, 1-9.

Luz, S., Haider, F., de la Fuente, S., Fromm, D., & MacWhinney, B. (2021). Detecting cognitive decline using speech only: The ADReSSo Challenge. *arXiv preprint arXiv:2104.09356*.

MacPherson, M. K. (2019). Cognitive load affects speech motor performance differently in older and younger adults. *Journal of Speech, Language, and Hearing Research, 62*(5), 1258-1277.

MacPherson, M. K., Abur, D., & Stepp, C. E. (2017). Acoustic measures of voice and physiologic measures of autonomic arousal during speech as a function

of cognitive load. *Journal of Voice, 31*(4), 504. e501-504. e509.

MacWhinney, B. (2019). Understanding spoken language through TalkBank. *Behavior research methods, 51*(4), 1919-1927.

March, E. G., Wales, R., & Pattison, P. (2006). The uses of nouns and deixis in discourse production in Alzheimer's disease. *Journal of Neurolinguistics, 19*(4), 311-340.

Mueller, P. B. (1997). *The aging voice.* Paper presented at the Seminars in speech and language.

Pappagari, R., Cho, J., Joshi, S., Moro-Velázquez, L., Zelasko, P., Villalba, J., & Dehak, N. (2021). *Automatic detection and assessment of Alzheimer Disease using speech and language technologies in low-resource scenarios.* Paper presented at the Proc. Interspeech.

Pastoriza-Domínguez, P., Torre, I. G., Diéguez-Vide, F., Gómez-Ruiz, I., Geladó, S., Bello-López, J., . . . Hernández-Fernández, A. (2022). Speech pause distribution as an early marker for Alzheimer's disease. *Speech Communication, 136*, 107-117.

Pistono, A., Jucla, M., Barbeau, E. J., Saint-Aubert, L., Lemesle, B., Calvet, B., . . . Pariente, J. (2016). Pauses during autobiographical discourse reflect episodic memory processes in early Alzheimer's disease. *Journal of Alzheimer's Disease, 50*(3), 687-698.

Pope, C., & Davis, B. H. (2011). Finding a balance: The carolinas conversation collection.

Ripich, D. N., Carpenter, B. D., & Ziol, E. W. (2000). Conversational cohesion patterns in men and women with Alzheimer's disease: a longitudinal study. *International journal of language & communication disorders, 35*(1), 49-64.

Sajjadi, S. A., Patterson, K., Tomek, M., & Nestor, P. J. (2012). Abnormalities of connected speech in semantic dementia vs Alzheimer's disease. *Aphasiology, 26*(6), 847-866.

Sankaranarayanan, S., Balan, J., Walsh, J. R., Wu, Y., Minnich, S., Piazza, A., . . . Bates, K. L. (2021). Covid-19 mortality prediction from deep learning in a large multistate electronic health record and laboratory information system data set: Algorithm development and validation. *Journal of medical Internet research, 23*(9), e30157.

Slegers, A., Filiou, R.-P., Montembeault, M., & Brambati, S. M. (2018). Connected speech features from picture description in Alzheimer's disease: A systematic review. *Journal of Alzheimer's Disease, 65*(2), 519-542.

Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., & Pakaski, M. (2015). Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. *Frontiers in aging neuroscience, 7*, 195.

Tabe-Bordbar, S., Emad, A., Zhao, S. D., & Sinha, S. (2018). A closer look at cross-validation for assessing the accuracy of gene regulatory networks and models. *Scientific reports, 8*(1), 1-11.

Tiwari, V. (2010). MFCC and its applications in speaker recognition. *International journal on emerging technologies, 1*(1), 19-22.

Tremblay, P., Deschamps, I., Bédard, P., Tessier, M.-H., Carrier, M., & Thibeault, M. (2018). Aging of speech production, from articulatory accuracy to motor timing. *Psychology and aging, 33*(7), 1022.

Trevethan, R. (2017). Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. *Frontiers in public health,*

5, 307.

Valsaraj, A., Madala, I., Garg, N., & Baths, V. (2021). *Alzheimer's Dementia Detection Using Acoustic & Linguistic Features and Pre-trained BERT.* Paper presented at the 2021 8th International Conference on Soft Computing & Machine Intelligence (ISCMI).

Wallin, A., Nordlund, A., Jonsson, M., Lind, K., Edman, Å., Göthlin, M., . . . Börjesson-Hanson, A. (2016). The Gothenburg MCI study: design and distribution of Alzheimer's disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up. *Journal of Cerebral Blood Flow & Metabolism, 36*(1), 114-131.

Wechsler, D. (1997a). *WAIS-3., WMS-3: Wechsler adult intelligence scale, Wechsler memory scale: Technical manual*: Psychological Corporation.

Wechsler, D. (1997b). *WMS-III: Wechsler memory scale administration and scoring manual*: Psychological Corporation.

Whitfield, J. A., Holdosh, S. R., Kriegel, Z., Sullivan, L. E., & Fullenkamp, A. M. (2021). Tracking the costs of clear and loud speech: Interactions between speech motor control and concurrent visuomotor tracking. *Journal of Speech, Language, and Hearing Research, 64*(6S), 2182-2195.

WHO. (2021). *Global status report on the public health response to dementia* . Retrieved from Geneva:

Xue, W., Cucchiarini, C., van Hout, R., & Strik, H. (2019). Acoustic correlates of speech intelligibility. The usability of the eGeMAPS feature set for atypical speech.

Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z., & Church, K. (2020). *Disfluencies and Fine-Tuning Pre-Trained Language Models for Detection of Alzheimer's*

Disease. Paper presented at the INTERSPEECH.

김하나, & 성지은. (2014). 노화에 따른 이야기 다시 말하기 수행력 및 작업 기억과의 상관관계 연구. [Age-related Changes in Story Retelling Procedures and their Relation to Working Memory Capacity]. *특수교육, 13*, 7-24. doi:10.18541/ser.2014.10.13.3.7

Supplementary Material 1. The Interview Questions

1) 올해 연세가 어떻게 되세요? 생년월일이 어떻게 되세요?

2) 학교는 어디까지 다니셨어요? 졸업은 하셨나요?

3) 어제 저녁 식사드셨죠? 몇시에, 누구와, 무엇을 드셨어요?

4) 저녁 드시고 TV도 보셨어요? 기억나는 장면 있으세요?

5) 요즘 기분 어떠세요?

Supplementary Material 2. The Modified Well-known Fairy-tale

Kongji and Patji (콩쥐팥쥐전)
어느 마을에 가난한 팥쥐와 팥쥐 엄마가 살고 있었습니다. 사람들은 더러운 팥쥐와 팥쥐 엄마를 마을에서 쫓아내려고 했습니다. 착한 콩쥐는 팥쥐와 팥쥐 엄마를 집에 데리고 와 함께 살았습니다. 원님은 콩쥐가 착한 일을 했다는 소문을 듣고 콩쥐를 신부로 맞이하였습니다.
The tale of Sim Cheong (심청전)
어느 마을에 심청이와 눈 먼 아버지가 함께 살고 있었습니다. 찢어지게 가난했던 심청이는 아버지를 바다에 밀어버렸습니다. 뒤늦게 죄책감에 엉엉 울던 심청이는 용왕님께 진심을 다해 용서를 빌었습니다. 심청이의 진심을 느낀 용왕님은 아버지를 다시 살려주었고 둘은 행복하게 살았습니다.

국 문 초 록

발화과제 성능비교: 자동화된 발화 기반 알츠하이머병 탐지를 위한 최적의 발화과제 탐색

서울대학교 인문대학

협동과정 인지과학 전공

배 민 주

연구목적: 음성은 최근 주목받고 있는 진단 마커 중 하나로 알츠하이머병의 조기진단을 용이하게 한다. 음성을 활용하여 알츠하이머병을 진단하고자 하는 그간의 선행연구들은 대개 알고리즘의 개선을 통한 분류 및 예측 성능의 향상을 도모하였으나 음성데이터 수집 단계에서 어떠한 발화과제가 알츠하이머병 발화의 특성을 효과적으로 유도 및 반영할 수 있을지에 대한 연구는 비교적 이루어지지 않았다. 본 연구에서는 참가자에게 반응 억제의 형태로 인지적 부하를 부과하는 새로운 발화과제를 제안하며 여러 발화과제의 분류 및 예측 성능을

비교하여 자동화된 발화기반 알츠하이머병 탐지도구로서의 가능성을 진단해보고자 한다.

연구방법: 본 연구에서는 79명의 알츠하이머병 환자와 79명의 정상노인에게 인터뷰, 따라말하기, 회상 과제를 수행하도록 하여 음성데이터를 수집하였다. 인터뷰 과제는 참가자의 일상생활에 관한 다섯 가지 질문으로 구성되었다. 따라말하기 과제와 회상 과제의 경우 두개의 수정된 유명 전래동화를 활용하여 진행되었다. 따라말하기 과제에서는 수정된 유명 전래동화를 구절 별로 들려주고 참가자가 이를 따라말하도록 하였다. 회상 과제에서는 원래 알고 있던 전래동화의 인출을 억제하며 들려준 수정된 유명 전래동화에서 획득한 새로운 정보를 최대한 자세하게 회상하도록 하였다. 음성데이터는 단일발화 단위로 분절하여 사용하였다. 인터뷰, 따라말하기, 회상 과제를 사용하여 수집된 발화데이터를 활용하여 각각 알츠하이머병 분류 모형과 인지손상 예측 모형을 만들었다. 모형 학습에 사용된 음성 특성들은 분산분석을 통해 통계적으로 유의미한 것으로 판단되었을 경우 사용되었다. 알츠하이머병 분류 모형에서는 랜덤 포레스트, 서포트 벡터 머신, 나이브 베이즈, k-근접이웃 기법을 사용하였으며 인지손상 예측 모형에서는 랜덤 포레스트, 서포트 벡터 머신, 릿지 기법을 사용하였다.

연구결과: 알츠하이머병 분류에서 가장 우수한 성능을 보인 모형은 72.9%의 교차검증 정확도를 보인 회상 과제 데이터셋으로 훈련된 랜덤포레스

트 모형이었다. 회상 과제 데이터셋으로 학습된 모형은 어떤 분류 알고리즘이 사용되었는가와 무관하게 다른 발화과제 데이터셋보다 우수한 성능을 보고하였다. 인지손상 예측에서 가장 우수한 성능을 보인 모형 역시 5.34 교차검증 RMSE와 4.38 교차검증 MAE를 보고한 회상 과제 데이터셋으로 훈련된 서포트 벡터 머신 모형이었다. 마찬가지로 어떤 예측 알고리즘의 경우에도 회상 과제에서 수집된 발화 데이터셋을 사용하였을 때 다른 발화과제에서 수집된 데이터셋을 통해 학습한 모형보다 더 우수한 예측 정확도를 보임을 알 수 있었다.

연구결론: 본 연구에서는 발화 데이터 수집 시 어떠한 발화과제를 사용하여 데이터를 수집하였는지가 알츠하이머병 분류와 인지손상 예측 모형의 성능에 영향을 미칠 수 있음을 확인하였다. 본 연구에서 사용한 인터뷰, 따라말하기, 회상 과제 중 회상 과제가 다른 발화과제에 비해 우수함을 보이는 것을 관찰되었다. 본 연구에서는 회상 과제 수행 시 참가자에게 부과된 인지적 부하가 발화 과정에 영향을 미쳐 알츠하이머병 환자의 특이적인 발화 특성이 보다 뚜렷하게 드러났을 가능성을 제안한다. 본 연구는 추후 연구에서 효과적인 자동화된 발화기반 알츠하이머병 탐지 도구를 개발하기 위하여 알츠하이머병 환자의 발화 특성을 풍부하게 반영할 수 있는 최적의 발화과제를 탐색 및 개발하여야 할 시사점을 제시한다.

주요어: 알츠하이머병, 간이정신상태검사, 음향음성학, 지도학습 머신러

닝

학　　번 : 2020-27753