The Relationship Between Elderly Speech Features in a Life-Logging Application and the Early Detection of Dementia

Naoka MATSUMURA[1], Yuta IZUTSU[1], Nobuyoshi KOMURO[2,*], Natsuko ARIMATSU[3], Aya MATSUMURA[3], Kota TOYAMA[1], Norimichi TSUMURA[1], Mizuki UMEHARA[3], and Ayumi AMEMIYA[3,*]

[1]Graduate School of Science and Engineering, Chiba University, Chiba, JAPAN
[2]Chiba University Digital Transformation Enhancement, Chiba University, Chiba, JAPAN
[3]Graduate School of Nursing, Chiba University, Chiba, JAPAN
[*]Corresponding author

***Abstract*—In this study, we aim to extract features from voice data of elderly individuals obtained through a life-logging application and evaluate mild cognitive impairment (MCI). Using the life-logging application, we collected approximately one month of voice data from 24 elderly participants (12 married couples, all aged 65 and over), and extracted features for the classification of MCI. As a result, sparse modeling outperformed other models in classification accuracy, achieving a score of 0.73. Furthermore, silent periods, shimmer, and filler duration were found to be important features for distinguishing MCI.**

***Index Terms*—Dementia, Early detection of dementia, Elderly people, Life-logging application,**

With the advancement of an aging society, it is estimated that approximately 50 million people worldwide are currently affected by Alzheimer's disease or dementia, and this number is expected to double every 20 years [1]. Dementia has become one of the major challenges in an aging society. Early detection of dementia offers significant benefits, including preparing caregiving environments, enhancing understanding of the condition, and slowing its progression.

Diagnosing dementia typically requires medical visits and examinations, which can be burdensome for both healthcare providers and elderly individuals. According to the Global Deterioration Scale, mild cognitive impairment (MCI) is defined as meeting two or more of the following seven criteria: difficulty remembering the date and time when visiting unfamiliar places, problems performing work tasks, difficulty recalling words or names, difficulty remembering sentences, inability to remember the names of introduced people, losing items, and decreased concentration [2].

One example of research aimed at the early detection of dementia is the use of blood biomarkers [3]. Biomarkers are objective indicators of disease status, obtained from body fluids or imaging tests. Blood tests can detect certain proteins that are considered to be associated with the development of Alzheimer's disease. Although cognitive function tests such as the MMSE (Mini-Mental State Examination) and MoCA-J (Japanese version of the Montreal Cognitive Assessment) are available to objectively evaluate cognitive function through scoring, it remains difficult to distinguish between mild cognitive impairment (MCI) and normal age-related decline. Additionally, examinees may try to mask their symptoms, making it hard to detect early signs.

Furthermore, when symptoms of dementia are identified, caregivers, who are family members of dementia patients, also experience a significant burden [4]. A negative correlation has been reported between health status and caregiving burden, with a correlation coefficient of -0.54 and a p-value less than 0.001. In addition to the overall caregiving burden, negative correlations were also observed for objective burden, stress-related burden, and interpersonal burden. Specifically, the correlation coefficients were -0.65 ($p < 0.001$) for health versus objective burden, -0.41 ($p = 0.001$) for health versus stress-related burden, and -0.29 for health versus interpersonal burden.

Currently, attempts are being made to reduce caregiver burden through the intervention of robots in caregiving settings. In some studies, voices recorded during MMSE tests and conversations with humanoid robots were analyzed, and the differences in voice features between healthy participants and MCI patients were examined [5]. The results indicated significant differences in speech duration, response time, silence duration, and voice fluctuation.

To reduce the burden on the elderly, a life-logging application has been developed. One of the early symptoms of dementia is emotional instability and reduced facial expressiveness [6]. One study using a life-logging application analyzed recorded data to infer the mood of elderly participants based on their facial expressions. The study classified mood into three categories—negative, neutral, and positive—and examined the relationship between facial expression features and mood. In addition, this application records users' responses to simple questions and collects voice data [7].

In the early stages of cognitive decline, mood fluctuations and difficulties in speech or understanding conversations may appear. Therefore, to objectively assess cognitive decline, it is necessary to examine the relationship between voice characteristics and cognitive function.

In this study, we aim to extract features from elderly individuals' voice data obtained through the life-logging application and evaluate mild cognitive impairment.

I. Proposed System

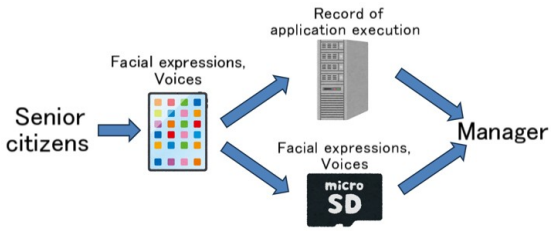


Fig. 1: System Overview

First, we introduce the system and operational aspects of the life-logging application. An overview of the system is

presented in Fig. 1. This application, developed using Unity for Android devices, is designed to record both facial expressions and voice. Users are only required to respond to daily questions presented via voice and text. Basic user information, as well as the start and end dates of the logging period, can be set by the administrator before the device is handed over to the user. Each day the user answers a question, a mark appears on a calendar within the app, allowing users to visualize their activity history and maintain motivation. Once the application begins asking a question, it starts recording, and the user's facial expression, voice, and response are saved. Voice is recorded and saved using Unity's built-in functionality. However, since Unity does not provide built-in video recording for facial expressions—only still images—OpenCV's VideoWriter is used to convert a series of still images into video, which is then saved. Audio and video data are initially saved in .wav and .avi formats, respectively. Due to their large file sizes, these are not suitable for long-term storage, so they are subsequently compressed and saved in .mp3 and .mp4 formats. Recorded data is stored in the local storage of the device. Whether the user has answered the day's questions is stored in the cloud using Firebase Realtime Database. Administrators can access the cloud from an administrator device to review user activity logs, enabling them to customize future questions based on each user's response frequency.

A total of 14 types of acoustic features were extracted: fundamental frequency, intensity (loudness), chroma features, mel-frequency cepstral coefficients (MFCC), utterance duration, speech rate, silent duration, silence rate, filler duration, filler rate, zero-crossing rate, response time, jitter, and shimmer. An overview of each feature is presented in Table I.

Since the Life-Logging Application covers approximately one month per participant, we computed the average, maximum, minimum, and variance for each feature per individual. We then examined the relationships between these features and the MMSE and MoCA-J scores, as well as correlations among the features themselves.

In this study, MCI was defined as an MMSE score of 27 or lower and a MoCA-J score of 25 or lower, and hypothesis testing was conducted for 13 healthy participants and 11 participants with MCI. Based on the hypothesis tests, the t-statistic, p-value, mean difference, and 95% confidence interval were calculated for each feature to examine whether there were significant differences.

Next, we applied machine learning techniques to classify individuals with cognitive decline based on the extracted acoustic features. For classification, the groups were defined in the same manner as in the hypothesis testing: participants with an MMSE score $\leqq$ 27 and a MoCA-J score $\leqq$ 25 were classified as MCI, yielding 13 healthy participants and 11 participants with MCI. Four classification algorithms were used: k-nearest neighbors, support vector machine (SVM), random forest, and sparse modeling. We investigated which algorithm achieved the highest classification accuracy. In addition, for the sparse modeling approach, we analyzed the importance of each acoustic feature in contributing to the classification.

The KNN method determines the class of a new data point by majority vote or average value. When a new data point is given, KNN refers to the K nearest neighbors. For classification problems, the data point is assigned to the class that occurs most frequently among the K neighbors. For regression problems, the average value of the K neighbors is used as the predicted value. In this study, classification is performed using majority vote. The advantages of KNN include immediate processing of new data and adaptability to high-dimensional or nonlinear data. However, disadvantages include high computational cost for large datasets and reduced effectiveness in very high-dimensional spaces. In this study, K=5 is used.

SVM seeks a hyperplane that maximally separates data points of different classes. A larger margin allows higher classification accuracy for new data. SVM can be linear or nonlinear. Linear SVM is used when the data can be separated by a straight line or plane. Nonlinear SVM is used when the data cannot be linearly separated, employing kernel methods to enable linear separation in a transformed feature space. SVM is effective for high-dimensional data and helps prevent overfitting but can be computationally expensive, similar to KNN. In this study, linear SVM is used with C=1.0 and random_state=42. Here, C is the regularization parameter, and random_state fixes the seed for random number generation.

Random Forest combines multiple decision trees for prediction. Each decision tree is constructed using a random subset of features, preventing over-reliance on specific features. For classification, the final result is determined by majority vote across all trees. Random Forest provides high accuracy and identifies important features but is not suitable for real-time processing. In this study, the parameters are n_estimators=100 and random_state=42.

Sparse modeling suppresses the influence of unnecessary features and selects only important features. This includes Lasso regression, which uses L1 regularization to set irrelevant feature coefficients to zero; compressed sensing used in signal and image processing; sparse principal component analysis with L1 regularization; and sparse representation for high-dimensional data. In this study, Lasso regression is used with random_state=42 and $\alpha = 0.01$, where α controls the degree of sparsity.

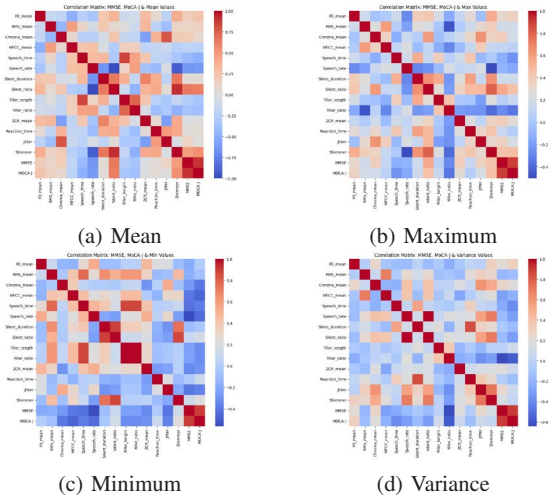
Twenty percent of the data was used as the test set and eighty percent as the training set to perform a single classification. At the same time, cross-validation was conducted. In cross-validation, the data was split into five folds, with one fold used as the test set and the remaining four folds used as the training set. This procedure was repeated five times, i.e., a 5-fold cross-validation was performed.

TABLE I: Extracted Acoustic Features

Feature Name	Description
Fundamental frequency	The lowest frequency component in the speech signal
Intensity (loudness)	The amplitude of the sound
Chroma features	Representation of the signal's frequency in 12 semitone bins
MFCC	Coefficients obtained by applying the discrete Fourier transform to a mel spectrum
Utterance duration	Duration of voiced segments
Speech rate	Ratio of utterance duration to total response time
Silent duration	Duration of unvoiced (silent) segments
Silence rate	Ratio of silent duration to total response time
Filler duration	Duration of filler words (e.g., "um", "uh") during speech
Filler rate	Ratio of filler duration to total response time
Zero-crossing rate	Number of times the signal's amplitude crosses the zero axis
Response time	Time between the end of the question and the start of the response
Jitter	Irregularity in pitch
Shimmer	Irregularity in amplitude

II. Results

The correlation matrix between each acoustic feature and the MMSE and MoCA-J scores is shown in Figure 1.



(a) Mean (b) Maximum

(c) Minimum (d) Variance

Fig. 2: Correlation matrices between each feature and cognitive test scores

First, positive correlations were observed between MMSE and MoCA-J scores for all measures, including mean, maximum, minimum, and variance. This is likely because higher scores on each test indicate normal cognitive function, whereas lower scores suggest the possibility of dementia-related symptoms. For the average values of each feature, positive correlations were found between jitter and chroma features, filler duration and speech duration, and shimmer and silent rate. In contrast, a negative correlation was observed between speech rate and shimmer. For the maximum values, in addition to this negative correlation, a negative correlation also appeared between filler rate and intensity. For the minimum values, a strong positive correlation was observed between filler duration and filler rate, while a negative correlation emerged between speech rate and MFCC.

The results of the hypothesis tests conducted for each feature are shown in Table II.

TABLE II: Results of hypothesis testing for each feature

feature	t_stat	p_value	mean_diff	ci_lower	ci_upper
F0_mean	-0.0206	0.9838	-0.3594	-34.5429	33.8241
RMS_mean	0.3973	0.6951	0.0007	-0.0027	0.0040
Chroma_mean	0.4661	0.6473	0.0116	-0.0372	0.0604
MFCC_mean	-0.2139	0.8327	-0.2558	-2.5998	2.0882
Speech_time	-2.0128	0.0702	-1.2253	-2.4185	-0.0321
Speech_rate	-2.1297	0.0524	-0.1380	-0.2650	-0.0110
Silent_duration	-0.4989	0.6270	-0.5177	-2.5516	1.5162
Silent_ratio	2.1297	0.0524	0.1380	0.0110	0.2650
Filler_length	-0.8100	0.4308	-0.0636	-0.2175	0.0903
Filler_ratio	-0.2875	0.7764	-0.0024	-0.0187	0.0139
ZCR_mean	-0.1308	0.8974	-0.0009	-0.0138	0.0121
Reaction_time	0.4267	0.6751	0.0502	-0.1805	0.2809
Jitter	-0.3281	0.7473	-0.1717	-1.1972	0.8539
Shimmer	2.0156	0.0587	0.2864	0.0079	0.5650

In all features, none of the p-values fell below 0.05. However, Speech_rate, Speech_time, Shimmer, and Silent_ratio, although exceeding 0.05, were below 0.1, indicating a tendency toward significance. We then examined the mean differences of these features showing such tendencies. Speech_rate had a negative mean difference, while Silent_ratio had a positive one. This suggests that individuals with MCI tend to spend less time speaking and exhibit longer silent periods. It can be inferred that language processing becomes more difficult, leading to hesitations or pauses during speech. Similarly, Speech_time, like Speech_rate, showed a negative mean difference, implying that individuals with MCI may find it difficult to sustain speech and often stop mid-sentence. On the other hand, Shimmer had a positive mean difference, suggesting that individuals with MCI have more difficulty maintaining stable vocal intensity compared to healthy controls. Regarding the confidence intervals, both the upper and lower bounds of Speech_time and Speech_rate were negative, while both bounds of Silent_ratio and Shimmer were positive. Although the sample size in this study was small and no feature reached complete statistical significance, these results suggest that temporal speech-related features and certain voice-quality measures may be associated with the detection of MCI.

Next, the classification reports for each learning method are shown in Tables IIIa to IIId.

TABLE III: Classification results of different methods

(a) k-NN

	precision	recall	f1-score	support
0	0.58	0.88	0.70	8
1	0.67	0.29	0.40	7
accuracy			0.60	15
macro avg	0.62	0.58	0.55	15
weighted avg	0.62	0.60	0.56	15

(b) SVM

	precision	recall	f1-score	support
0	0.83	0.62	0.71	8
1	0.67	0.86	0.75	7
accuracy			0.73	15
macro avg	0.75	0.74	0.73	15
weighted avg	0.76	0.73	0.73	15

(c) Random Forest

	precision	recall	f1-score	support
0	0.75	0.75	0.75	8
1	0.71	0.71	0.71	7
accuracy			0.73	15
macro avg	0.73	0.73	0.73	15
weighted avg	0.73	0.73	0.73	15

(d) Sparse Modeling (Lasso)

	precision	recall	f1-score	support
0	0.83	0.62	0.71	8
1	0.67	0.86	0.75	7
accuracy			0.73	15
macro avg	0.75	0.74	0.73	15
weighted avg	0.76	0.73	0.73	15

Looking at the results of KNN shown in Table IIIa, the accuracy was 0.60. While the recall for healthy controls was 0.88, indicating that they were classified relatively accurately, the recall for individuals with MCI was only 0.29, showing that they were barely identified.

In contrast, for SVM, Random Forest, and Sparse Modeling, the accuracy was 0.73, slightly higher than that of KNN. Moreover, for both healthy controls and individuals with MCI, recall remained at least 0.6 or higher. These three methods were able to classify more accurately than KNN, even with the limited sample size.

Table IV shows the features with the highest importance when classification was performed using Random Forest.

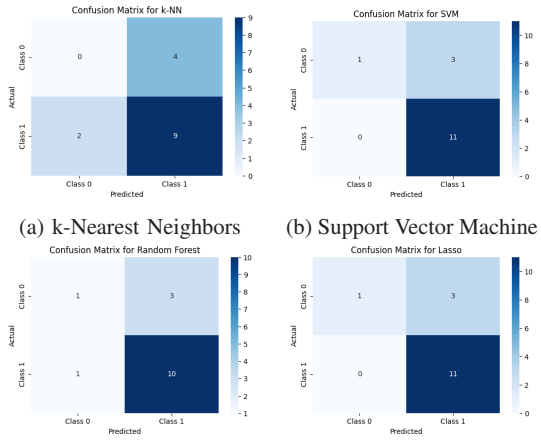
Since all importance scores were below 10%, it can be inferred that the classification was not driven by a small

TABLE IV: Feature Importances

Feature	Importance
Shimmer	0.0933
Speech_rate	0.0926
ZCR_mean	0.0913
Jitter	0.0847
F0_mean	0.0768
Speech_time	0.0717
RMS_mean	0.0709
Silent_ratio	0.0704
Chroma_mean	0.070

number of dominant features, but rather by the combined contribution of multiple features. Among the features with high importance were Shimmer and Jitter, indicating that individuals with MCI tend to have less stability in their voice. In addition, consistent with the results of the hypothesis testing, Speech_rate, Silent_ratio, and Speech_time also ranked among the top features in terms of importance.

The following figures show the classification results of the k-nearest neighbors method (Figure 3a), SVM (Figure 3b), Random Forest (Figure 3c), and sparse modeling (Figure 4).



(a) k-Nearest Neighbors (b) Support Vector Machine

(c) Random Forest (d) Sparse Modeling

Fig. 3: Classification results for each algorithm

Although individuals exhibiting dementia symptoms were correctly classified using k-nearest neighbors, Random Forest, and Sparse Modeling, the classification accuracy for cognitively healthy individuals was relatively lower. Figure 4 shows the feature importance derived from Sparse Modeling.

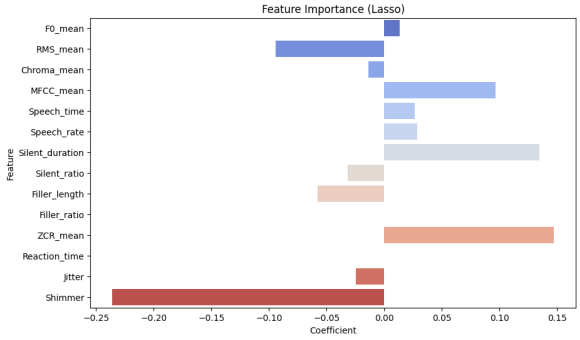


Fig. 4: Regression coefficients of each feature

In the classification using Sparse Modeling, Shimmer contributed the most in the negative direction among all features. In addition to Shimmer, intensity also contributed negatively, while zero-crossing rate and silent duration contributed positively.

As a result of cross-validation, the accuracy of the k-nearest neighbors method was 0.4552 ± 0.1325, that of SVM was 0.7467 ± 0.1431, that of Random Forest was 0.4829 ± 0.1679, and that of Sparse Modeling was 0.6505 ± 0.1054. In the single validation test, SVM, Random Forest, and Sparse Modeling showed the same accuracy; however, after performing cross-validation, SVM achieved the highest accuracy. Nevertheless, the standard deviations were relatively large across all models constructed in this study.

III. Conclusion

Using speech data from the life-logging application, we extracted features to evaluate their relationship with dementia symptoms.

By applying SVM and sparse modeling, relatively accurate classification was achieved using only the recorded "a i u e o" utterances from a tablet device. Additionally, both hypothesis testing and machine learning analyses showed that features such as volume and fundamental frequency were not strongly associated with dementia symptoms. However, it was observed that voice stability measures, such as Shimmer and Jitter, tended to be unstable, and participants with MCI exhibited frequent pauses and longer silent periods while thinking. Although participants with dementia were correctly classified, the classification accuracy for healthy participants was low. Therefore, future work should focus on developing methods that reduce false positives for healthy individuals. Furthermore, since the current study included only 24 participants, the models were unstable; thus, classification using a larger sample size is necessary.

References

[1] A. Warren, "An integrative approach to dementia care," Frontiers in Aging, 2023, doi: 10.3389/fragi.2023.1143408.

[2] K. Toba, "Early detection of dementia in elderly patients from a clinical perspective," in Proc. 48th Annual Meeting of the Japanese Geriatrics Society, Symposium II: Early Detection and Treatment of Dementia, 2007

[3] E. Ausó, V. Gómez-Vicente, G. Esquiva, "Biomarkers for Alzheimer's Disease Early Diagnosis," Journal of Personalized Medicine, 10(3):114, 2020

[4] C. O. Bailes, C. M. Kelley, and N. M. Parker, "Caregiver burden and perceived health competence when caring for family members diagnosed with Alzheimer's disease and related dementia," Journal of the American Association of Nurse Practitioners, vol. 28, no. 10, pp. 534–540, 2016.

[5] K. Yoshii, D. Kimura, S. Kosugi, K. Shinkawa, T. Takase, M. Kobayashi, Y. Yamada, M. Nemoto, R. Watanabe, E. Tsukada, M. Ota, J. Higashi, K. Nemoto, T. Arai, and M. Nishimura, "Preliminary study for simple dementia screening using daily conversation speech with a humanoid robot," *IPSJ SIG Technical Report*, vol. 2020-SLP-133, no. 7, 2010.

[6] N. Arimatsu, A. Matsumura, Y. Tahara, Y. Izutsu, A. Kawasaki, K. Toyama, N. Komuro, N. Tsumura, and A. Amemiya, "Relationship between mood and facial expression in daily life of community-dwelling elderly people," 2024.

[7] Y. Izutsu, N. Komuro, N. Arimatsu, A. Matsumura, K. Toyama, N. Tsumura, Y. Tahara, and A. Amemiya, "Development of a life-logging application using facial expressions, voices, and actions for the early detection of dementia and caregiver mental fatigue," 2024.