Dementia Prediction From Speech Signal Using Optimized Prosodic Features

Bagus Tris Atmaja*, Sakriani Sakti*
* Nara Institute of Science and Technology, Japan
E-mail: bagus.tris@naist.ac.jp, ssakti@is.naist.jp

***Abstract*—Early detection of dementia, a general term for cognitive decline, is vital for enhancing patient care and management. Conventional clinical assessments and cognitive tests require expert clinicians and are time-consuming. Therefore, developing rapid and reliable in-house detection methods is highly desirable. Prior studies have identified speech-based early markers of dementia, motivating the application of machine learning for automated detection. This study investigates the effectiveness of optimized prosodic features, with a focus on pause-related metrics, for dementia classification. Experimental evaluation of 15 selected features shows that pause and formant features are the most discriminative, achieving improved classification performance over the original 45-feature set. These findings suggest that concise, targeted feature sets can enhance automated speech-based dementia detection.**

I. Introduction

Recent studies have explored acoustic features for detecting dementia from speech. Important acoustic features include duration-related measures, prosodic features, and paralinguistic characteristics [1], [2]. The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) has shown promise in dementia detection. Combining acoustic features with linguistic features, such as BERT-based models, can further improve classification accuracy. Multi-Resolution Cochleagram (MRCG) features and active data representation (ADR) methods have also been introduced for dementia recognition [2]. Feature selection techniques have identified both acoustic and linguistic characteristics as valuable for dementia screening, with a combination of seven types of characteristics producing the best results [3]. These studies demonstrate that acoustic features, especially when combined with linguistic features, can effectively contribute to automated dementia detection from speech.

Specific speech features, such as speech rate and pause duration, have been determined as potential markers for detecting dementia. Research has shown that these temporal speech parameters can effectively differentiate between healthy controls, mild cognitive impairment (MCI), and Alzheimer's disease (AD) patients [4]–[7]. Specifically, speech rate decreases while the number and duration of pauses increase as cognitive impairment progresses [4], [5]. The proportion of pauses to total speech time also grows with dementia severity [8]. Automated analysis of speech samples has demonstrated a progressive increase in pause duration from healthy controls to mild and moderate dementia groups [6]. Furthermore, the probability density distribution of the durations of the pauses follows a lognormal distribution, with AD patients showing longer pauses and greater variability compared to healthy controls and MCI patients [7].

This paper contributes to implementing the previously correlated speech features for dementia into machine learning. We extended the prosodic features with new pause-based features and evaluated these features in different numbers based on their importance. The new optimized prosodic features showed improvement over the raw prosodic features in two datasets and a combination of both.

II. Methods

A. Dataset and Data Splitting

We evaluated two datasets and a mix of both, which is described below.

DementiaNet[1] contains public figures with a confirmed dementia diagnosis and other public figures with no cognitive decline (control) over the age of eighty. The data were gathered from YouTube, where public figures are interviewed with spontaneous speech.

Positive audio samples were obtained from public figures and celebrities with confirmed diagnoses of dementia. Following verification of each individual's diagnosis, the author of dataset collected interview recordings from multiple time periods relative to the onset or confirmation of symptoms: (i) within 0–5 years post-diagnosis, (ii) 5–10 years prior to diagnosis, and (iii) 10–15 years prior to diagnosis. This temporal stratification was aimed at capturing speech characteristics potentially associated with the progression of dementia.

To minimize the risk of data leakage and false negatives, stringent inclusion criteria were applied. The authors of the dataset selected public figures and celebrities with strong evidence of cognitive health into advanced age. Specifically, individuals included were either confirmed centenarians, currently living and aged 85 or older, or deceased at age 90 or above with no recorded history or signs of dementia. For each individual, three audio samples were collected from distinct age brackets: (i) after age 70, (ii) between ages 55 and 70, and (iii) before age 55. This sampling approach was designed to ensure age diversity while maintaining confidence in the cognitive health status of the subjects.

The original samples of DementiaNet are 227 for training and 48 for validation. We allocated samples for validation for

[1]https://github.com/shreyasgite/dementianet

979-8-3315-7206-8/25/$31.00

the test set since the machine learning method (XGBoost) is utilized. From 227 samples in the training set, 106 samples are dementia, and the other 121 samples are controls. For the test set, 20 samples are dementia and 28 samples are controls. The total number of unique speakers is 141. All samples are provided in WAV format.

Demenetiabank (Pitt Corpus) is a shared database of multimedia interactions for the study of communication in dementia [9]. The original dataset contains 117 people diagnosed with Alzheimer's Disease, and 93 healthy people, reading a description of an image, and the task is to classify these groups. The release used in this evaluation contains only the audio part of this dataset, without the text features. The version is 'Pitt $>$ Cookie' in The Talkbank System (not `Pitt-orig` or `0extra`). We evaluated the original MP3 version of the audio files.

We used the split version for training, validation, and test from Tensorflow Dataset[2]. We follow this split for ease and for a consistent benchmark for future work. The total number of audio files is 552 samples, with 243 samples for control and 309 samples for dementia. The number of unique speakers is 292.

Perfect speaker separation is achieved for the training and test splits. There are 350 samples for training (validation files are merged into training) and 102 samples for the test (50 dementia + 52 control). The number of dementia samples in training is 259, and the number of control samples is 191.

A mixed dataset is a concatenation of DementiaNet with DementiaBank. Training samples from DementiaNet are merged with training samples from DementiaBank, as well as for the test split. On all three datasets, either no balancing, balancing with ros [10], or balancing with SMOTE [10] is used, and the method that yields the highest accuracy score is reported.

B. Speech Features

We extracted speech features from the Praat toolkit via parselmouth [11], [12]. The origin of those speech features could be traced from the Praat script to detect syllable nuclei and measure speech rate automatically [13]. We extended the number of features from 39 to 45 from the previous publication [14]. The new additions include the following features:

- pause_lognorm_mu - Location parameter (μ) of the lognormal distribution fitted to pause durations
- pause_lognorm_sigma - Shape parameter (σ) of the lognormal distribution fitted to pause durations
- pause_lognorm_ks_pvalue - P-value from Kolmogorov-Smirnov test for goodness of fit of the lognormal distribution
- pause_mean_duration - Mean duration of pauses between speech segments
- pause_std_duration - Standard deviation of pause durations
- pause_cv - Coefficient of variation (CV) of pause durations (std/mean)
- proportion_pause_duration - Proportion of total pause duration relative to speaking time

Note that the previous version of Nkululeko [15], the toolkit used to extract speech features and do experimentation, includes two 'duration' variables with different names, of which one is removed in the current version. Hence, adding seven new features resulted in 45 features. A list of all 45 features is shown in Table I.

TABLE I
LIST OF 45 RAW SPEECH FEATURES DERIVED FROM PRAAT TOOLKIT

'duration', 'meanF0Hz', 'stdevF0Hz', 'HNR', 'localJitter', 'localabsoluteJitter', 'rapJitter', 'ppq5Jitter', 'ddpJitter', 'localShimmer', 'localdbShimmer', 'apq3Shimmer', 'apq5Shimmer', 'apq11Shimmer', 'ddaShimmer', 'f1_mean', 'f2_mean', 'f3_mean', 'f4_mean', 'f1_median', 'f2_median', 'f3_median', 'f4_median', 'pause_lognorm_mu', 'pause_lognorm_sigma', 'pause_lognorm_ks_pvalue', 'pause_mean_duration', 'pause_std_duration', 'pause_cv', 'nsyll', 'npause', 'phonationtime_s', 'speechrate_nsyll_dur', 'articulation_rate_nsyll_phonationtime', 'ASD_speakingtime_nsyll', 'proportion_pause_duration', 'JitterPCA', 'ShimmerPCA', 'pF', 'fdisp', 'avgFormant', 'mff', 'fitch_vtl', 'delta_f', 'vtl_delta_f'

C. Models

The XGB or XGBoost (e**X**treme **G**radient **B**oosting) [16] is implemented in Nkululeko, following the standard gradient boosting framework. XGBoost is an ensemble learning method that combines multiple weak learners (decision trees) to create a robust predictive model.

XGBoost minimizes the following objective function:

$$L(\varphi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$$

Where:

- $l(y_i, \hat{y}_i)$ is the loss function between true labels y_i and predictions $\hat{y}_i$;
- $\Omega(f_k)$ is the regularization term for the k-th tree;
- $\varphi = f_1, f_2, ..., f_k$ represents the ensemble of K trees.

We used the default parameters provided by Nkululeko for running the experiments.

For optimized prosodic features, we evaluated XGB and SHAP [17]. We reported which one of these models provides higher accuracy for the same number of 15 features. These 15 features will be fed again to the XGB model to compare the performance with the baseline of 45 features.

III. RESULTS AND DISCUSSION

A. Results of Individual and Mixed Datasets

We found that feature selection with SHAP performed better for DementiaNet, while for DementiaBank, feature selection via the 15 most important features from the XGB model yielded higher accuracy than SHAP analysis. The list of features is listed in Figure 1 and Table II.

The SHAP (SHapley Additive exPlanations) feature importance analysis presented in Figure 1 reveals critical insights

[2]https://www.tensorflow.org/datasets/catalog/dementiabank

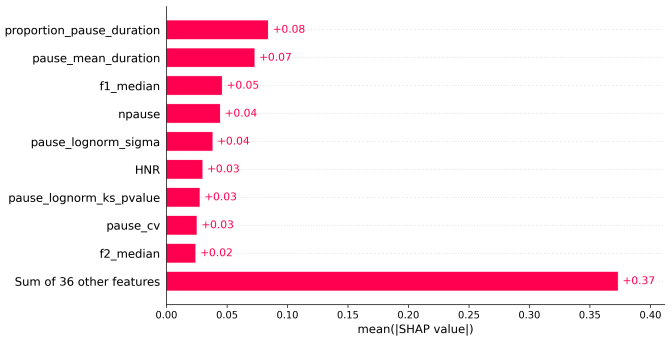

Fig. 1. Top 9 speech features related to DementiaNet Dataset

into the acoustic biomarkers most predictive of dementia classification using an XGBoost model trained on the DementiaNet dataset. The analysis employs 45 carefully engineered acoustic features extracted through Praat-based analysis, representing a comprehensive characterization of speech pathology indicators.

The results demonstrate a clear hierarchy of feature importance, with **pause-related temporal features dominating the top rankings**. The most discriminative feature is 'proportion_pause_duration' (importance: 0.084), which quantifies the ratio of pause time to speaking time and represents a fundamental disruption in speech fluency characteristic of cognitive decline. This is followed by 'pause_mean_duration' (0.073) and 'f1_median' (0.046), indicating that both temporal patterning and articulatory precision are compromised in dementia. Notably, **six of the top ten features relate to pause distribution analysis**, including novel lognormal distribution parameters ('pause_lognorm_sigma', 'pause_lognorm_kspvalue') that capture the statistical properties of pause timing irregularities. The prominence of formant frequencies (F1, F2 medians) and voice quality measures (HNR: 0.030, ShimmerPCA: 0.023) suggests that both articulatory control and vocal fold dynamics serve as complementary biomarkers. Traditional jitter and shimmer measures occupy lower importance rankings, with several showing zero contribution (ddpJitter, ddaShimmer), indicating that the enhanced pause distribution features provide superior discriminative power for dementia detection compared to conventional voice perturbation measures.

Table II presents the feature importance rankings derived from XGBoost classification applied to the DementiaBank dataset, revealing distinct patterns in acoustic biomarker significance compared to other dementia corpora. The analysis demonstrates the overwhelming dominance of the third formant median frequency (f3_median) with an importance score of 0.07530, representing approximately 84% of the total feature contribution and suggesting that high-frequency articulatory control may be particularly compromised in the DementiaBank population. This exceptional prominence of f3_median, which relates to tongue tip and blade positioning during speech production, indicates that fine motor control of anterior articulation serves as a primary diagnostic indicator in this dataset. The substantial gap between the top-ranked feature and subsequent parameters (pause_lognorm_ks_pvalue: 0.00803, pause_lognorm_mu: 0.00550) underscores the critical role of formant-based assessment in distinguishing cognitive impairment within the DementiaBank paradigm.

The feature hierarchy reveals complementary contributions from temporal and prosodic parameters, with pause distribution modeling (pause_lognorm_ks_pvalue, pause_lognorm_mu) and basic speech metrics (npause: 0.00537, duration: 0.00417) occupying intermediate importance levels. Notably, the relatively modest contribution of proportion_pause_duration (0.00236) in this dataset contrasts with its higher ranking in other corpora, suggesting dataset-specific variations in the manifestation of temporal speech disruptions. Traditional voice quality measures, including jitter parameters (localabsoluteJitter: 0.00248, ppq5Jitter: 0.00115) and fundamental frequency characteristics (meanF0Hz: 0.00066), demonstrate diminished importance compared to articulatory features, reinforcing the hypothesis that DementiaBank tasks may be particularly sensitive to formant-based degradation patterns. This feature distribution pattern indicates that automated detection systems for the DementiaBank dataset should prioritize sophisticated formant analysis while incorporating pause distribution statistics as secondary diagnostic indicators.

TABLE II
TOP 15 FEATURES RELATED TO DEMENTIABANK DATASET

Feature	XGB Importance
f3_median	0.07530
pause_lognorm_ks_pvalue	0.00803
pause_lognorm_mu	0.00550
npause	0.00537
duration	0.00417
nsyll	0.00290
f1_mean	0.00278
localabsoluteJitter	0.00248
proportion_pause_duration	0.00236
ppq5Jitter	0.00115
meanF0Hz	0.00066
ASD_speakingtime_nsyll	0.00042
pF	0.00042
apq3Shimmer	0.00030
pause_mean_duration	0.00012

Table III presents the ranked feature importance scores derived from the XGBoost classifier for the mixed dataset, revealing the relative contribution of acoustic parameters to dementia classification performance. The analysis demonstrates that formant frequency measures constitute the most discriminative features, with the first formant median frequency (f1_median) achieving the highest importance score of 0.02894, followed by the third formant median (f3_median) at 0.00915. The prominence of formant-based features aligns with established phonetic theory, as formant frequencies directly reflect vocal tract configuration and articulatory precision, both of which are compromised in neurodegenerative conditions. The substantial contribution of proportion_pause_duration (0.01306) further underscores the significance of temporal speech characteristics in dementia detection, consistent with clinical observations

of increased hesitations and speech planning difficulties in cognitively impaired populations.

The order of importance also reveals the effect of advanced pause distribution modeling, with pause_lognorm_ks_pvalue (0.00867) and pause_lognorm_mu (0.00242) demonstrating that statistical characterization of pause patterns provides complementary diagnostic information beyond simple temporal measures. Fundamental frequency parameters, including meanF0Hz (0.00697) and stdevF0Hz (0.00516), occupy intermediate positions in the ranking, reflecting the documented relationship between prosodic control and cognitive decline. Notably, traditional voice quality measures such as harmonics-to-noise ratio (HNR: 0.00161) and shimmer-based parameters (localdbShimmer: 0.00343) exhibit relatively lower importance scores, suggesting that articulatory and temporal features may be more sensitive indicators of dementia-related speech changes than conventional voice perturbation measures. This feature importance distribution supports a multi-dimensional approach to acoustic biomarker development, emphasizing the complementary roles of formant analysis, temporal patterning, and prosodic characteristics in automated dementia detection systems.

TABLE III
FEATURE IMPORTANCE RANKINGS FROM XGBOOST MODEL FOR MIXED DATASET

Feature	XGB Importance
f1_median	0.02894
proportion_pause_duration	0.01306
f3_median	0.00915
pause_lognorm_ks_pvalue	0.00867
meanF0Hz	0.00697
ASD_speakingtime_nsyll	0.00605
f1_mean	0.00540
stdevF0Hz	0.00516
npause	0.00387
pF	0.00359
localdbShimmer	0.00343
f3_mean	0.00294
pause_lognorm_mu	0.00242
f4_mean	0.00198
HNR	0.00161

Table V presents a comprehensive evaluation of classification performance across varying feature dimensionalities, demonstrating the efficacy of feature reduction strategies in dementia detection tasks. The analysis reveals that DementiaNet consistently achieves superior performance across all feature reduction scenarios, with unweighted accuracy (UA) ranging from 0.729 to 0.760, weighted accuracy (WA) from 0.742 to 0.770, and F1-scores from 0.786 to 0.807. Notably, the DementiaNet dataset exhibits remarkable resilience to dimensionality reduction, maintaining competitive performance even with minimal feature sets (3 features: UA=0.750, WA=0.742, F1=0.786), suggesting that the acoustic characteristics captured in this corpus are highly discriminative and that the SHAP-based feature selection methodology effectively identifies the most informative parameters. The consistently high performance with reduced feature sets indicates that DementiaNet's speech tasks may elicit more pronounced acoustic biomarkers or that the dataset's demographic composition facilitates more robust feature-based discrimination.

In contrast, DementiaBank demonstrates greater sensitivity to feature reduction, with performance degrading from UA=0.659 (10 features) to UA=0.617 (3-5 features), while maintaining relatively stable F1-scores between 0.668-0.702. The mixed dataset configuration exhibits an intermediate performance profile, achieving optimal results with five features (UA=0.710, WA=0.700, F1=0.731) before experiencing a minor degradation at three features. The fact that XGBoost-based feature selection dominates the DementiaBank and mixed dataset scenarios, while SHAP methodology proves optimal for DementiaNet, suggests dataset-specific optimization requirements for feature selection algorithms. These findings collectively demonstrate that effective dementia detection can be achieved with minimal feature sets ($\leq$10 features), supporting the development of computationally efficient clinical screening tools while highlighting the importance of dataset-appropriate feature selection strategies.

The comparative analysis presented in Table IV demonstrates significant performance improvements achieved through optimized feature selection methodologies relative to baseline configurations across three experimental datasets. The baseline system, employing all 45 available acoustic features, establishes reference performance levels with DementiaNet achieving the highest discriminative capacity (UA=0.746, WA=0.771, F1=0.820), followed by DementiaBank (UA=0.669, WA=0.666, F1=0.707), and the mixed dataset exhibiting the most challenging classification scenario (UA=0.611, WA=0.601, F1=0.619). The optimized configurations, utilizing targeted feature selection strategies with only 15 carefully chosen parameters, demonstrate substantial performance gains across all evaluation metrics, with DementiaNet experiencing improvements of +0.050 in UA, +0.041 in WA, and +0.027 in F1-score when employing SHAP-based feature selection. Similarly, DementiaBank achieves notable enhancement (+0.079 UA, +0.079 WA, +0.069 F1) through XGBoost-derived feature importance rankings, while the mixed dataset configuration realizes the most dramatic improvement (+0.110 UA, +0.119 WA, +0.093 F1).

These results underscore the critical importance of feature engineering and selection in acoustic-based dementia detection systems, demonstrating that dimensionality reduction from 45 to 15 features not only enhances computational efficiency but also improves classification performance by eliminating redundant or noisy parameters. The consistently superior performance of optimized configurations across all datasets validates the hypothesis that carefully curated minimal feature sets can outperform comprehensive feature collections, supporting the development of clinically viable screening tools that balance diagnostic accuracy with computational tractability and interpretability requirements. The UMAP visualization in Fig. 2 also shows the separation of dementia vs. control labels using 15 features, although there is a mix of both labels mainly in the healthy label area.

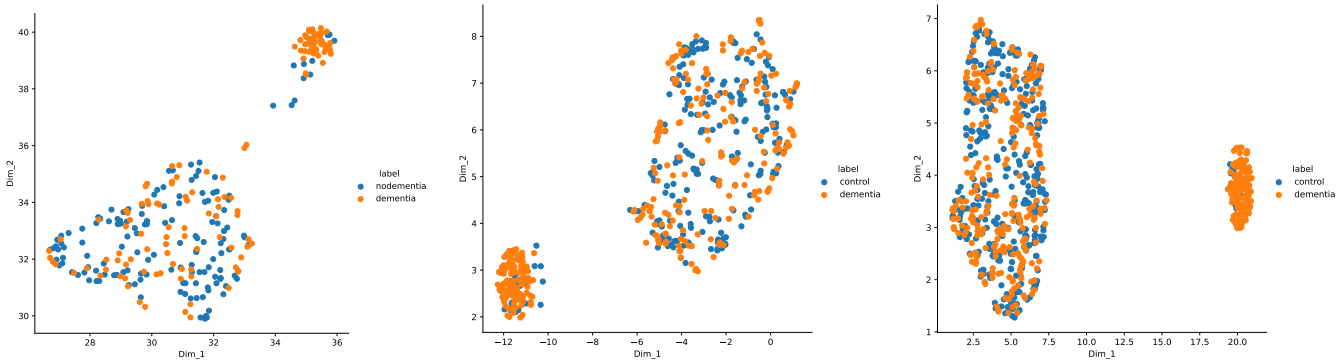

Fig. 2. Plots of UMAP from best (15) features for each dataset (Left: DementiaNet, Middle: DementiaBank, Right: Mixed dataset)

TABLE IV
PERFORMANCE OF BASELINE AND OPTIMIZED FEATURES

Dataset	UA	WA	F1-score	n_feat-model
		Baseline		
DementiaNet	0.746	0.771	0.820	45
DementiaBank	0.669	0.666	0.707	45
Mixed dataset	0.611	0.601	0.619	45
		Optimized		
DementiaNet	0.796	0.812	0.847	15-shap
DementiaBank	0.748	0.745	0.776	15-xgb
Mixed dataset	0.721	0.720	0.712	15-xgb

TABLE V
OPTIMAL PERFORMANCE WITH MINIMUM FEATURES

Dataset	UA	WA	F1	Model
		10 features		
DementiaNet	0.760	0.770	0.807	SHAP
DementiaBank	0.659	0.656	0.701	XGB
Mix dataset	0.683	0.673	0.707	XGB
		5 features		
DementiaNet	0.729	0.750	0.800	SHAP
DementiaBank	0.617	0.622	0.668	XGB
Mixed dataset	0.710	0.700	0.731	XGB
		3 features		
DementiaNet	0.750	0.742	0.786	SHAP
DementiaBank	0.617	0.623	0.702	XGB
Mixed dataset	0.694	0.680	0.727	XGB

B. *Effect of Reducing Number of Features*

We previously found that the performance of the mixed dataset is between DementiaNet and DementiaBank, meaning that the performance can be predicted linearly. This phenomenon is shown strongly when reducing the number of features. Compared to 15 features, the performance of evaluation using 10, 5, and 3 features is lower than that of 15 features. The exception is that for a mixed dataset, the performance of 5 and 3 features is higher than that of 10 features. Also, for DementiaNet and DementiaBank, the performance of 5 features provides a balanced trade-off between 10 and 5 features. The speech features that appear in the intersection of all 10 features are 'f1_mean', 'npause', 'pause_lognorm_ks_pvalue', and 'proportion_pause_duration'. However, the performance of those five features is lower when evaluated on each dataset compared to the five features optimized on each dataset.

C. *Benchmark*

Table VI presents a comparative performance evaluation of different acoustic feature extraction methods for dementia detection tasks, with particular emphasis on feature dimensionality and classification accuracy across multiple datasets. The benchmark compares several established feature extraction approaches against the proposed method implemented in this study. The proposed method demonstrates several notable achievements: (1) achieves competitive or superior performance using only 15 features, significantly fewer than most baseline approaches (except PRAT with seven features), indicating effective feature selection and engineering; (2) the method shows robust performance across two distinct datasets; (3) on-par performance: on the Pitt Corpus, the proposed method outperforms IS10 features while using significantly fewer dimensions (15 per utterance vs. 75 per frame), though MFCC++ achieves higher accuracy with 44 features; (4) balanced classification: The close alignment between UA and WA scores (particularly evident in DementiaNet results: 0.796 vs. 0.812) suggests relatively balanced class distributions and consistent performance across demographic groups.

The superior performance on DementiaNet (UA=0.796, WA=0.812, F1=0.847) compared to Pitt Corpus suggests that the feature set may be particularly well-suited for the acoustic characteristics present in this dataset. The high F1-score of 0.847 indicates strong discriminative capability for dementia detection, which is crucial for clinical screening applications where both sensitivity and specificity are paramount.

The computational efficiency achieved through the reduced feature dimensionality (n_feat=15) represents a significant advantage for real-world deployment, potentially enabling real-time processing and reducing computational overhead in clinical settings while maintaining diagnostic accuracy.

IV. CONCLUSIONS

This study presents an evaluation of optimized prosodic features for dementia detection. We extended the previous speech features extracted from the Praat toolkit with seven new pause-based speech features. Analysis of feature importance via SHAP and XGB models reveals that most of those features

TABLE VI
BENCHMARK OF DIFFERENT METHODS WITH DIFFERENT NUMBER OF FEATURES

Method	n_feat	Dataset	UA	WA	F1-score
PRAT [18]	7	Elderly	-	0.582	0.621
IS10 [19]	75	Pitt Corpus	-	0.731	-
MFCC++ [20]	44	Pitt Corpus	-	0.876	0.875
eGeMAPS [21]	88	ADReSSo	0.730	0.746	0.750
This study	15	Pitt Corpus	0.748	0.745	0.776
This study	15	DementiaNet	0.796	0.812	0.847

are ranked top among other formant features. We then selected 15 features to be compared with the original 45 features for dementia detection from speech. Results showed that 15 features consistently achieved higher accuracy for DementiaNet, DementiaBank, and the mixed dataset (UA=0.75, 0.80, and 0.72, respectively). This small number of features is beneficial for real-time applications.

ACKNOWLEDGMENT

This paper is partly based on results obtained from projects JSPS KAKENHI Grant Number 24K0296. Pitt corpus was supported by the National Institutes of Health grants NIA AG03705 and AG05133.

REFERENCES

[1] N. Liu, Z. Yuan, and Q. Tang, "Improving Alzheimer's Disease Detection for Speech Based on Feature Purification Network," *Front. Public Heal.*, vol. 9, no. March, pp. 1–9, 2022, ISSN: 22962565.

[2] F. Haider, S. de la Fuente, and S. Luz, "An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer's Dementia in Spontaneous Speech," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 2, pp. 272–281, Feb. 2020, ISSN: 1932-4553.

[3] J. Weiner and T. Schultz, "Detection of intra-personal development of cognitive impairment from conversational speech," in *Speech Communication; 12. ITG Symposium*, VDE, 2016, pp. 1–5.

[4] E. Gonzalez-Moreira, D. Torres-Boza, M. A. Garcia-Zamora, C. A. Ferrer, and L. A. Hernandez-Gomez, "Prosodic Speech Analysis to Identify Mild Cognitive Impairment," in *VI Latin American Congress on Biomedical Engineering CLAIB 2014, Paraná, Argentina 29, 30 & 31 October 2014*, 2015, pp. 580–583.

[5] V. Vincze, G. Szatlóczki, L. Tóth, *et al.*, "Telltale silence: temporal speech parameters discriminate between prodromal dementia and mild Alzheimer's disease," *Clin. Linguist. Phon.*, vol. 35, no. 8, pp. 727–742, Aug. 2021, ISSN: 0269-9206.

[6] R. A. Sluis, D. Angus, J. Wiles, *et al.*, "An Automated Approach to Examining Pausing in the Speech of People With Dementia," *Am. J. Alzheimer's Dis. Other Dementias®*, vol. 35, Jan. 2020, ISSN: 1533-3175.

[7] P. Pastoriza-Domínguez, I. G. Torre, F. Diéguez-Vide, *et al.*, *Speech pause distribution as an early marker for Alzheimer's disease*, Jan. 2021.

[8] R. Haulcy and J. Glass, "Clac: A speech corpus of healthy english speakers," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 1, 2021, pp. 201–205, ISBN: 9781713836902.

[9] J. T. Becker, "The Natural History of Alzheimer's Disease," *Arch. Neurol.*, vol. 51, no. 6, p. 585, Jun. 1994, ISSN: 0003-9942.

[10] G. Lemaitre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, Sep. 2016, ISSN: 15337928. eprint: 1609.06570.

[11] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *J. Phon.*, vol. 71, no. 2018, pp. 1–15, 2018, ISSN: 00954470.

[12] D. R. Feinberg, *Parselmouth Praat Scripts in Python*, 2018.

[13] N. H. de Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behav. Res. Methods*, vol. 41, no. 2, pp. 385–390, 2009, ISSN: 1554351X.

[14] B. T. Atmaja and A. Sasou, "Pathological Voice Detection From Sustained Vowels : Handcrafted vs. Self-supervised Learning," in *2025 IEEE Int. Conf. Acoust. Speech, Signal Process. Work.*, 2025.

[15] F. Burkhardt, B. T. Atmaja, A. Derington, and F. Eyben, "Check Your Audio Data : Nkululeko for Bias Detection," in *Orient. COCOSDA*, IEEE, Oct. 2024, pp. 1–6, ISBN: 979-8-3315-0603-2.

[16] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., Curran Associates, Inc., 2017, pp. 4765–4774.

[18] K. Nishikawa, H. Kawano, R. Hirakawa, and Y. Nakatoh, "Analysis of Prosodic Features and Formant of Dementia Speech for Machine Learning," in *Proc. - 2022 5th Int. Conf. Inf. Comput. Technol. ICICT 2022*, 2022, pp. 173–176, ISBN: 9781665469609.

[19] M. Rodrigues Makiuchi, T. Warnita, N. Inoue, *et al.*, "Speech paralinguistic approach for detecting dementia using gated convolutional neural network," *IEICE Trans. Inf. Syst.*, vol. 104, no. 11, pp. 1930–1940, 2021, ISSN: 17451361. eprint: 2004.07992.

[20] M. R. Kumar, S. Vekkot, S. Lalitha, *et al.*, "Dementia Detection from Speech Using Machine Learning and Deep Learning Architectures," *Sensors*, vol. 22, no. 23, 2022, ISSN: 14248220.

[21] "Predicting dementia from spontaneous speech using large language models," *PLOS Digit. Heal.*, vol. 1, no. 12 December, pp. 1–14, 2022, ISSN: 27673170.