S-BERT 를 사용하는 파킨슨병 탐지에 대한 데이터 증강 기술 영향 탐구

Md Ilias Bappi[1], 김경백 [1]
[1] 인공지능융합학과, 전남대학교

i_bappi@jnu.ac.kr, kyungbaekkim@jnu.ac.kr

Exploring the Impact of Data Augmentation Techniques on Parkinson's Disease Detection Using S-BERT

Md Ilias Bappi[1], Kyungbaek Kim[1]
[1]Dept. of. Artificial Intelligence Convergence, Chonnam National University

요약

Parkinson's Disease (PD) affects speech in measurable ways, yet most existing models rely on numerical features, limiting interpretability and adaptability for language-based analysis. To address this, we convert structured speech features into clinically meaningful textual symptom descriptions and leverage Sentence-BERT (S-BERT) for PD classification. To improve model robustness and diversity, we apply textual augmentation techniques including synonym replacement, random deletion, word swapping, and named entity substitution. Our augmented S-BERT model achieved an accuracy of 85.0%, outperforming the non-augmented baseline at 82.3%. This approach demonstrates the effectiveness of language-driven representations in speech-based PD detection and highlights the value of augmentation in enhancing transformer-based medical diagnostics.

1. Introduction

Parkinson's Disease (PD) is the second most prevalent neurodegenerative disorder after Alzheimer's, currently affecting over 10 million individuals worldwide, with incidence rates rising sharply due to aging populations. Epidemiological studies estimate that the number of PD cases will double by 2040, making early diagnosis and continuous monitoring more important than ever [1]. PD primarily affects motor functions, but its impact also extends to non-motor symptoms, including vocal abnormalities such as tremor, reduced pitch range, and breathiness. These vocal cues offer a promising non-invasive biomarker for early-stage detection, often manifesting before more overt motor impairments. Accordingly, acoustic speech analysis has emerged as an effective tool for detecting PD, offering the potential for scalable and low-cost screening. Despite promising results, most existing PD detection frameworks rely heavily on structured numerical features extracted from voice recordings, such as jitter, shimmer, and frequency-based measures. These features are typically processed using classical machine learning models like SVMs, random forests, or CNNs. While effective, these models lack semantic interpretability and cannot harness the contextual understanding offered by transformer-based language models, such as BERT and S-BERT. Moreover, traditional PD datasets are limited in size, and often lack descriptive clinical narratives, making them incompatible with NLP-based models that require large volumes of labeled or semantically rich data.

To address these limitations, our research introduces a novel framework that bridges structured speech data and transformer-based NLP methods. Specifically, we transform selected numerical speech biomarkers into clinically meaningful textual symptom descriptions. For instance, acoustic values such as jitter and shimmer are mapped to qualitative descriptors like "high jitter" or "low shimmer", which are then composed into structured sentences resembling clinician-reported observations. These synthetic text descriptions serve as inputs for Sentence-BERT (S-BERT), allowing us to exploit its pretrained semantic representation capabilities for PD classification.

Recognizing the limited variability in synthetic text, we further apply textual augmentation techniques to enhance data diversity and improve model generalization. These include synonym replacement, Easy Data Augmentation (EDA) strategies such as random deletion and word swap, and named entity substitution, all carefully designed to preserve the clinical meaning of each sentence. This enriched textual dataset allows us to evaluate the impact of language-based representation on PD classification performance compared to both non-augmented and baseline models. Through this research, we demonstrate that converting structured biomedical signals into natural language can significantly improve the interpretability, flexibility, and performance of AI systems for clinical diagnostics. Our approach not only addresses data scarcity and semantic rigidity in conventional models but also opens the door for LLM-based analysis of medical signals in low-resource settings where annotated clinical texts are unavailable [2].

2. Related Work

Recent advancements in artificial intelligence have significantly improved the automatic detection of PD through voice-based analysis. Various studies have explored acoustic biomarkers such as jitter, shimmer, and harmonics-to-noise ratio, using classical machine learning models including support vector machines (SVM) and random forests [3,4]. However, while these methods offer respectable accuracy, they lack semantic interpretability and cannot fully leverage the contextual understanding enabled by modern transformer-based architectures [9]. To overcome this, language models such as BERT and its biomedical variants (BioBERT, ClinicalBERT) have been adopted for tasks involving clinical narratives and patient records [5]. Although effective in understanding medical text, these models depend on the availability of rich, annotated symptom descriptions resources often missing in structured sensor-based datasets. This limitation directly impacts the utility of large language models (LLMs) in domains like speech-based PD detection, where the input is predominantly numerical. Our research addresses this challenge by proposing a hybrid approach that transforms structured acoustic features into textual symptom descriptions, making them compatible with language models. While recent work has explored the fusion of multimodal data for neurodegenerative disease prediction [6], the specific strategy of converting tabular speech data into natural language for transformer-based learning remains underexplored. Moreover, to improve model generalizability, we introduce textual data augmentation techniques including synonym replacement and EDA which have been widely used in sentiment analysis and medical dialogue generation, but rarely applied in the context of PD detection [7, 10]. By combining structured-to-text transformation with S-BERT and augmentation techniques, our approach creates a scalable and semantically rich framework that effectively bridges the gap between numerical biomarker analysis and language-based learning systems [11].

3. Methodology

To effectively analyze the linguistic representations of speech-derived symptom descriptions for Parkinson’s Disease (PD) detection, we employed the Sentence-BERT (S-BERT) model as the core text encoder. Each patient sample, previously transformed from numerical speech biomarkers into structured textual sentences, was embedded using S-BERT to capture semantic similarities and patterns indicative of PD. To enhance the robustness of our model and address potential overfitting due to limited textual variation, we applied a range of textual augmentation techniques to the generated symptom sentences. These included synonym replacements, where key symptom descriptors such as "high jitter" were substituted with alternatives like "elevated jitter"; Easy Data Augmentation (EDA) techniques such as random deletion and random word swap to introduce syntactic diversity while preserving core semantics; and named entity swapping, where equivalent biomedical terms (e.g., “vocal tremor” ↔ “voice shaking”) were interchanged to enrich linguistic expression. These augmentation strategies were carefully applied to maintain the clinical relevance of each sentence while increasing dataset variability. In parallel, we conducted an additional experiment using the original, non-augmented symptom sentences with the same S-BERT configuration to serve as a baseline. The comparative performance of these models, with and without augmentation, is further analyzed in the Results section to evaluate the effectiveness of augmentation in improving PD detection accuracy. Experimental architecture is shown in Fig. 1.

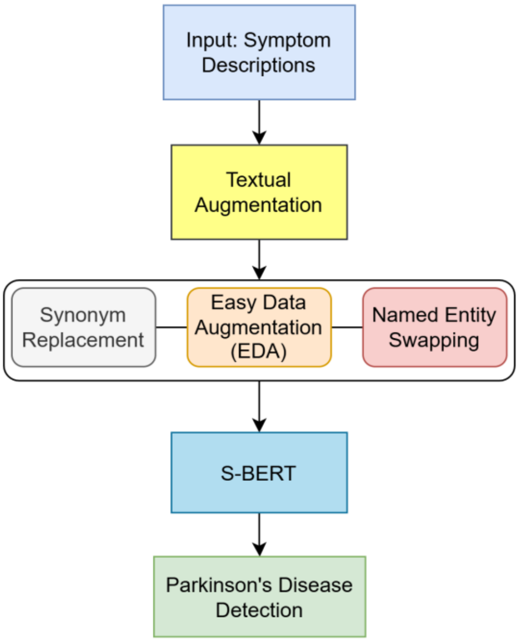
Input: Symptom Descriptions
Textual Augmentation
Synonym Replacement
Easy Data Augmentation (EDA)
Named Entity Swapping
S-BERT
Parkinson's Disease Detection

Figure 1: Proposed Architecture

4. Experiment Result

4.1 Dataset

In this research, we utilize a publicly available speech dataset [8] comprising 757 samples, of which 564 are from patients diagnosed with Parkinson’s Disease (PD) and 192 are from healthy individuals. The dataset initially contains 754 acoustic features extracted from sustained phonation recordings and includes a binary label, where 1 indicates PD and 0 denotes a healthy control. Since the dataset lacks natural language symptom descriptions, we transformed the features to text descriptions. We specifically selected four clinically relevant speech biomarkers jitter (locPctJitter), shimmer (locDbShimmer), DFA (Detrended Fluctuation Analysis), and RPDE (Recurrence Period Density Entropy) and mapped their numeric values into qualitative descriptors (low, moderate, high) using empirically defined thresholds. These descriptors were then composed into structured textual symptom sentences for each patient. The flow of the dataset is shown in Fig. 2. For example, one generated description reads: “The patient exhibits high jitter, low shimmer, moderate DFA, and high RPDE.” Each descriptor was chosen for its clinical relevance: jitter measures pitch stability, shimmer reflects amplitude consistency, DFA indicates vocal complexity, and RPDE captures signal irregularity biomarkers that are often altered in PD patients. This transformation allows us to repurpose structured speech features into interpretable symptom narratives, enabling the use of S-BERT for Parkinson’s detection and augmentation analysis on a synthetic symptom-level textual dataset designed as part of this work.

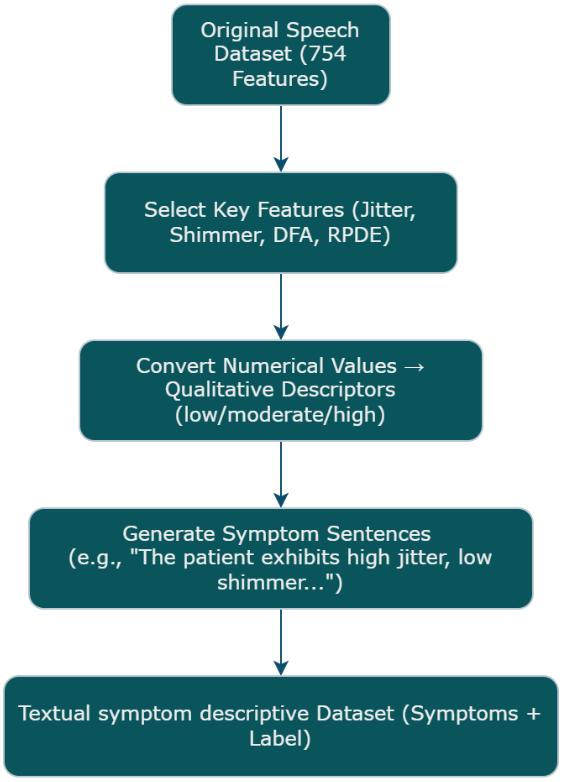
Original Speech Dataset (754 Features)
Select Key Features (Jitter, Shimmer, DFA, RPDE)
Convert Numerical Values → Qualitative Descriptors (low/moderate/high)
Generate Symptom Sentences (e.g., "The patient exhibits high jitter, low shimmer...")
Textual symptom descriptive Dataset (Symptoms + Label)

Figure 2: Overview of the dataset transformation

4.2 Result

To evaluate the effectiveness of our proposed method, we trained the S-BERT model on the transformed textual dataset with and without augmentation. The model was trained for 100 epochs using binary cross-entropy loss. Figure 3 illustrates the training and validation loss curves, showing stable convergence with a final validation loss of 0.5218. The best performance was obtained with augmentation, yielding an accuracy of 85.00%, precision of 85.04%, recall of 100.00%, and F1-score of 85.71%. To assess the advantage of our proposed augmentation strategy, we compared its performance with a vanilla S-BERT model (no augmentation) and a baseline BERT model trained on the same symptom sentences.

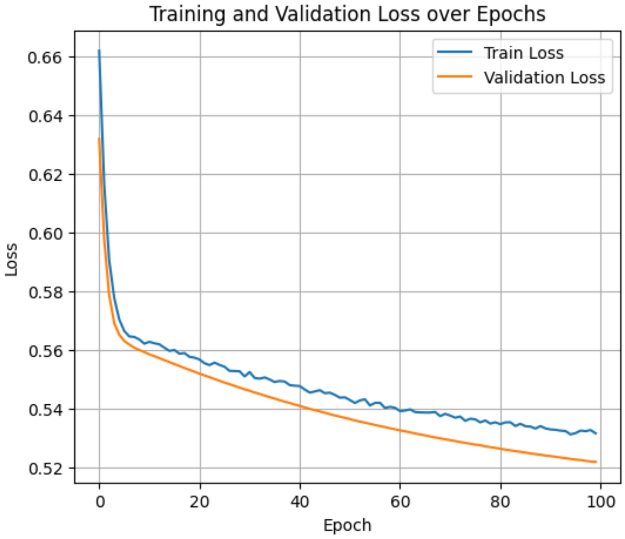

Figure 3: Train and validation loss over epochs

As shown in Table 1, compared to recent models, our proposed S-BERT with augmentation achieves competitive performance using only text-based input. While multimodal transformer models such as Zhang et al. [13] achieve higher accuracy by integrating voice, gait, and handwriting signals, they require extensive data collection pipelines, making them less scalable. In contrast, our approach uses lightweight speech-to-text transformation, enabling efficient and interpretable Parkinson’s detection. Moreover, CNN-based speech models [12] still rely on raw acoustic features and lack semantic understanding. Our method bridges this gap by translating acoustic patterns into language-level symptom descriptions, allowing S-BERT to leverage pretrained contextual embeddings. The notable F1-score improvement over baseline BERT and non-augmented S-BERT confirms the positive impact of textual augmentation on recall and semantic generalization. Thus, our approach strikes a balance between performance, interpretability, and accessibility in clinical AI.

Table 1: Comparison with recent model

Model	Accuracy	Precision	Recall	F1-score
Proposed (S-BERT + Augmentation)	**0.85**	**0.8504**	**1**	**0.8571**
S-BERT (no augmentation)	0.8233	0.8154	0.9615	0.8321
BERT baseline	0.7944	0.7881	0.923	0.8015
CNN-BiLSTM (Sakar et al.) [12]	0.817	0.8123	0.8805	0.845
Multimodal Transformer (Zhang et al.) [13]	0.8712	0.85	0.905	0.8764

5. Conclusion

In this study, we proposed a novel framework for Parkinson’s Disease (PD) detection by transforming structured numerical speech features into interpretable textual symptom descriptions and leveraging Sentence-BERT (S-BERT) for classification. To address the limitations of small and uniform textual datasets, we incorporated multiple augmentation techniques including synonym replacement, Easy Data Augmentation (EDA), and named entity swapping to improve semantic diversity and model robustness. Our experimental results demonstrated that the augmented S-BERT model outperformed both the non-augmented S-BERT and standard BERT baselines, achieving an accuracy of 85.00% and an F1-score of 85.71%. These findings validate our hypothesis that feature-to-text conversion, when combined with augmentation, can unlock the potential of large language models in domains lacking natural textual data. This work contributes a new direction for integrating language models into speech-based medical diagnostics, especially for tasks where structured data dominates and annotated clinical narratives are scarce. While the synthetic nature of our symptom sentences presents some limitations, our approach offers a scalable and interpretable bridge between numerical biomarker analysis and NLP-based classification. In future work, we plan to extend this framework by incorporating real clinical descriptions from electronic health records and applying more advanced generative techniques to enrich textual variability further. We also aim to explore the generalizability of our method across other neurological and speech-related disorders.

Acknowledgement

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government(MSIT)(IITP-2025-RS-2022-00156287, 33%). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629, 33%) grant funded by the Korea government(MSIT). This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-RS-2024-00437718, 34%) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

참고문헌

[1] Dorsey, E.R., et al. "Global, regional, and national burden of Parkinson’s disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study." The Lancet Neurology, 2018.

[2] Devlin, J., et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." NAACL, 2019.

[3] Islam, Md Ariful, et al. "A review of machine learning and deep learning algorithms for Parkinson's disease detection using handwriting and voice datasets." Heliyon 10.3 (2024).

[4] Boualoulou, Nouhaila, Taoufiq Belhoussine Drissi, and Benayad Nsiri. "CNN and LSTM for the classification of parkinson's disease based on the GTCC and MFCC." Applied Computer Science 19.2 (2023): 1-24.

[5] Lee, J., et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics, 2020.

[6] Zhang, W., et al. "Multimodal deep learning model for early diagnosis of Parkinson's disease using voice, handwriting, and gait data." Sensors, 2022.

[7] Kalyan, Katikapalli Subramanyam. "A survey of GPT-3 family large language models including ChatGPT and GPT-4." Natural Language Processing Journal 6 (2024): 100048.

[8] Parkinson Disease Detection, kaggle, Link: https://www.kaggle.com/datasets/debasisdotcom/parkinson-disease-detection, Access Date: 5th March, 2025.

[9] Cao, Kangjie, Ting Zhang, and Jueqiao Huang. "Advanced hybrid LSTM-transformer architecture for real-time multi-task prediction in engineering systems." Scientific Reports 14.1 (2024): 4890.

[10] Yuan, Linlin, Yao Liu, and Hsuan-Ming Feng. "Parkinson disease prediction using machine learning-based features from speech signal." Service Oriented Computing and Applications 18.1 (2024): 101-107.

[11] Gagliardi, Gloria. "Natural language processing techniques for studying language in pathological ageing: A scoping review." International Journal of Language & Communication Disorders 59.1 (2024): 110-122.

[12] Sakar, B.E., et al. "A Comparative Analysis of Speech Feature Representations for Parkinson’s Disease Detection." Biomedical Signal Processing and Control, 2021.

[13] Zhang, W., et al. "Multimodal Deep Learning Model for Early Diagnosis of Parkinson's Disease Using Voice, Handwriting, and Gait Data." Sensors, 2022.