

하이트노이즈 환경에서 모음과 파열음 기반 파킨슨병 분류 성능의 저하 및 견고성 평가

박현주¹, 정희용¹, 김경백¹
¹ 전남대학교 인공지능융합학과

{qkrguswn1114, h.jeong, kyungbaekkim}@jnu.ac.kr

Performance Degradation and Robustness Evaluation of Parkinson's Disease Classification Based on Vowels and Plosives under White-Noise Conditions

Hyeonju Park¹, Hieyong Jeong¹, Kyungbaek Kim¹
¹Dept. of AI Convergence, Chonnam National University

요 약

파킨슨병(Parkinson's disease, PD)의 조기 진단을 위해 음성 기반 분석 기법이 활발히 연구되고 있다. 본 연구에서는 MFCC 기반 음성 특징 이미지를 입력으로 하는 Vision Transformer(ViT) 모델에 Partial Class Activation Attention(PCAA) 기법을 적용하여, PD와 정상인(Healthy Control, HC)의 분류 성능 및 해석 가능성을 평가하였다. 두 개의 공개 데이터셋(IPVS, Voice Samples)을 통합하여 총 676 개(HC:305, PD:371)의 데이터를 사용하였으며, ResNet50 및 기본 ViT 모델과 성능을 비교하였다. 그 결과, 제안된 ViT+PCAA 모델은 8:1:1 분할 실험에서 91.18%의 정확도를 기록하였고, 잡음 환경(SNR 0~30 dB)에서도 전 구간에서 가장 우수한 성능을 보였다. 특히 0 dB 조건에서도 58% 정확도를 유지하며, 30 dB에서는 87%에 도달하여 잡음에 대한 강건성이 크게 향상됨을 확인하였다. 본 연구는 PCAA 기반 Attention 기법이 음성 기반 질환 분류의 성능과 설명력을 개선할 수 있음을 보여주며, 향후 임상 적용 가능성을 뒷받침한다.

1. 서론

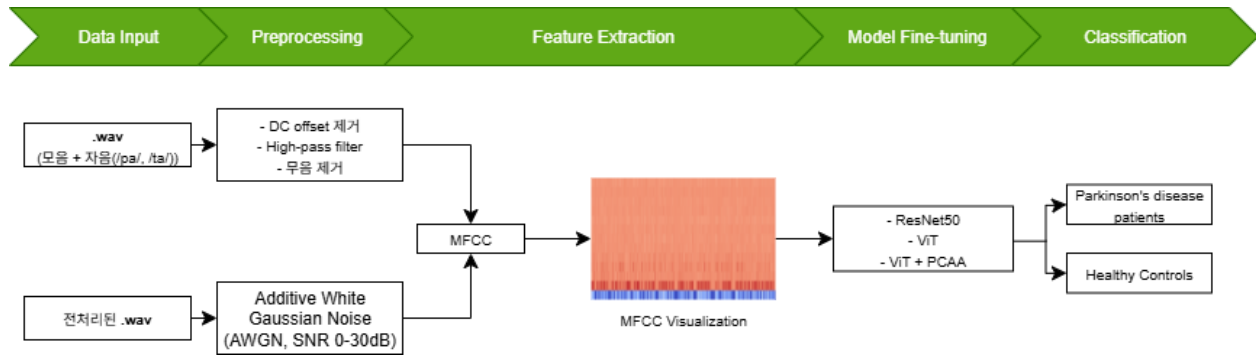
파킨슨병(Parkinson's disease, PD)은 운동 및 음성 장애를 유발하는 대표적인 신경퇴행성 질환이다. 최근 PD 환자와 정상인(Healthy control, HC)을 분류하기 위해 음성데이터를 활용한 연구가 계속 진행되고 있다. 특히 모음(/a/, /e/, /i/, /o/, /u/)과 파열음(/p/, /t/ 등)은 성대 및 발성 기관의 미세한 운동 변화를 반영하기 때문에 진단적 단서로서의 가치가 높다고 알려져 있다.[1] 기존 연구 대부분은 SVM, Random Forest 등 전통적인 분류기를 기반으로 특정 음향 특성의 조합[2] 또는 발화 유형 조합[1]이 PD 구분에 얼마나 효과적인지를 탐색하는 데 초점을 두고 있다. 그러나 이러한 접근은 모델의 판단 근거를 시각적으로 해석하는 데 한계가 있다.

본 연구에서는 전통적인 분류기 대신 Vision Transformer(ViT)[3]에 Partial Class Activation Attention (PCAA) 기법[4]을 결합한 모델을 제안한다. PCAA는 기존의 CAM 대비 이미지의 국소(부분) 수준

활성화를 정교하게 포착할 수 있다. 따라서 본 연구는 PCAA 기반 Attention 모델을 적용하여 MFCC 음성 특징 이미지를 입력으로 사용하여 성능과 해석 가능성을 분석한다. 또한 제안 모델의 분류 정확도를 다양한 기준 모델과 비교·분석하고, 잡음을 추가한 음성을 MFCC로 표현했을 때의 성능 변화를 함께 평가하여 임상 및 실사용 환경에서의 강건성 또한 검증하고자 한다.

2. 관련 연구

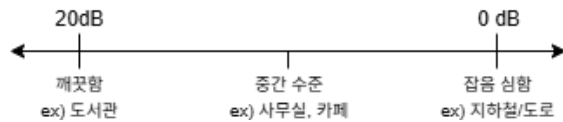
기존 PD 음성 분류 연구는 주로 MFCC와 기초 음향 지표를 결합하여 SVM, KNN과 같은 전통적인 분류기를 적용하는 방식으로 진행되었으며, 일부 연구에서는 98% 이상의 높은 정확도를 보고한 바 있다.[1,2] 다만 이러한 접근은 잡음 환경에서의 강건성 분석이 충분히 이루어지지 않았다는 한계가 있다. 최근에는 멜 스펙트로그램을 입력으로 활용하는 CNN 및 Transformer 기반 딥러닝 접근이 활용되고 있으며, 사전학습된 CNN을 미세조정해 PD 분류에서 AUC 0.92~0.97의 성능을 달성한 사례도 보고되었다.[5]



(그림 1) 연구 아키텍처 개요

더 나아가 음성 이외의 보행 시계열 데이터를 영상화하여 딥러닝으로 분석하는 연구도 제안되었다.[6]

한편, 실 환경에서의 음향 SNR 은 대체로 0~30dB 범위에 분포하며, 20dB 이하에서는 WER(Word Error Rate, 단어 오류율)이 급격히 증가하는 것으로 알려져 있다.[7,8] 그림 2 와 같이 SNR(dB)값을 실생활 소음 환경에 대응시켜 정리한 예시를 보여준다.



(그림 2) SNR(dB) 수준에 따른 실생활 소음 환경 예시

본 연구는 이러한 맥락 속에서 MFCC 기반 특징 이미지를 Vision Transformer(ViT)에 입력하고, PCAA 기법을 적용하여 성능을 개선하는 동시에, 다양한 SNR 조건에서 모델의 강건성을 검증하고자 한다.

3. 연구 방법

3.1 사용 모델 및 데이터셋

본 연구에서는 ImageNet-21K(약 1,400 만 장 이미지, 21,000 개 클래스)로 사전학습된 ViT 모델[9,10]을 기반으로 하여 일부 레이어(0, 5, 11)에 PCAA 기법을 적용하였고, Resnet50[11] 및 기본 ViT 와 비교하였다. 데이터는 두 가지 파킨슨병 음성 데이터셋을 활용하였다. 첫째, Italian Parkinson's Voice and Speech(IPVS) [12-14] 데이터셋으로, 다양한 발화 유형으로 구성되어 있으며, PD 28 명, HC 37 명의 음성이 포함된다. 본 연구는 실험의 일관성을 위해 약 5 초 길이의 지속 모음(/a/) 및 음절 발화(/pa/, /ta/, 각 5 초) 샘플만을 선별하였다. 둘째, Voice Samples for Patients with Parkinson's Disease and Healthy Controls[15] 데이터셋으로, PD 40 명과 HC 41 명의 “지속 모음 /a/” 발화가 포함되어 있다. 두 데이터셋을 통합한 결과, HC 집단에서는 305 개, PD 집단에서는 371 개의 데이터를 확보하였다.

3.2 오디오 전처리

모든 오디오는 전처리(DC 제거, 필터링, 무음 제거) 후 MFCC 특징(13 차원)을 추출하여 이미지 형태로 변환하고, 학습, 검증, 테스트를 8:1:1 로 분할하

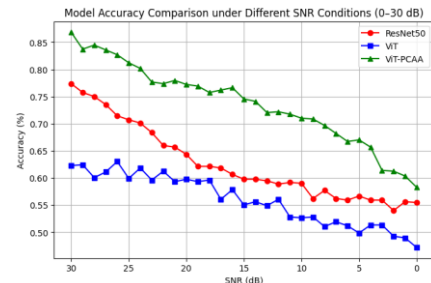
였다. 표 1 은 모델의 테스트 정확도를 비교한 결과이다. 이후 학습된 모델에 대해 AWGN 을 적용해 SNR 0~30dB 환경에서 성능 변화를 평가하였다.

<표 1> MFCC 기반 오디오 데이터를 활용한 8:1:1 분할 실험에서의 테스트 정확도

Model	Test Accuracy
ResNet50	94.12%
ViT	58.82%
ViT + PCAA	91.18%

4. 실험 결과 및 분석

기본 ViT 모델의 정확도가 58.82%로 상대적으로 낮게 나타난 이유는 본 연구의 데이터 규모가 제한적이기 때문으로 판단된다. ViT 는 self-attention 기반



(그림 3) 다양한 SNR 조건(0 ~ 30 dB)에서의 모델 정확도 비교

의 구조로, 대규모 데이터에서 전역적 패턴을 학습하는 데 강점을 보이지만, 데이터가 적을 경우 패치 간 상관관계 학습이 불안정해질 수 있다. 반면, PCAA 를 결합한 ViT+PCAA 는 부분적 활성화 정보를 활용하여 주요 특징 영역에 집중함으로써 학습 효율이 개선된 것으로 보인다. 이는 ViT 구조의 데이터 효율성을 보완하는 방향으로 해석할 수 있다.

그림 4 는 SNR(0~30dB) 조건에서 세 모델의 성능 변화를 나타낸다. 모든 모델은 SNR 증가에 따라 정확도가 향상되었으며, 그중 ViT+PCAA 가 전 구간에서 가장 높은 성능을 기록하였다. 특히 0dB 에서 58%, 30dB 에서 87%의 정확도를 달성하여, 잡음 환경에 대한 강건성이 기존 모델 대비 뚜렷하게 개선됨을 확인할 수 있었다.

5. 결론

본 연구에서는 MFCC 기반 음성 데이터를 입력으로 하여 파킨슨병 환자(PD)와 정상인(HC)을 분류하기 위한 Vision Transformer(ViT) 모델에 Partial Class Activation Attention(PCAA) 기법을 적용하였다. 제안된 ViT+PCAA 모델은 기존 ResNet50 및 기본 ViT 모델과 비교하여 전반적으로 더 높은 정확도와 잡음 환경에 대한 강건성을 보였다.

다만 본 연구는 총 676 개의 음성 샘플을 기반으로 수행되어, 데이터 규모 측면에서 일반화 가능성에 한계가 존재한다. 향후 연구에서는 다양한 언어, 발화 조건, 녹음 환경을 포함한 확장된 데이터셋을 활용하여 모델의 안정성과 재현성을 검증할 계획이다. 또한 임상 환경에서의 적용 가능성을 평가함으로써 실제 진단 지원 시스템으로의 발전 가능성을 탐색하고자 한다.

사 사

이 논문은 한국연구재단 기초연구사업 (No. RS-2021-NR066151)과 정보통신기획평가원 핵심전략 R&D 사업 (No. IITP-2025-RS-2025-02219190)의 지원을 받아 수행되었다.

참고문헌

- [1] Aishat Toye, A. et al., "Comparative Study of Speech Analysis Methods to Predict Parkinson's Disease", arXiv e-prints, Art. no. arXiv:2111.10207, 2021.
- [2] Scimeca, Sabrina et al., "Robust and language-independent acoustic features in Parkinson's disease." *Frontiers in neurology* vol. 14 1198058. 13 Jun. 2023.
- [3] Dosovitskiy Alexey, Beyer Lucas, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", arXiv e-prints, 2010.11929, 2020.
- [4] Sun-Ao Liu, Hongtao Xie et al., "Partial Class Activation Attention for Semantic Segmentation", 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, 16836-16845.
- [5] Rahmatallah, Yasir, et al. "Pre-trained convolutional neural networks identify Parkinson's disease from spectrogram images of voice samples." *Scientific Reports* 15.1.2025, 7337.
- [6] Setiawan, Febryan, and Che-Wei Lin., "Implementation of a deep learning algorithm based on vertical ground reaction force time-frequency features for the detection and severity classification of Parkinson's disease." *Sensors* 21.15.2021, 5207.
- [7] Smeds, Karolina, Florian Wolters et al., "Estimation of signal-to-noise ratios in realistic sound scenarios.", *Journal of the American Academy of Audiology*, 2015, 183-196.
- [8] SELTZER, Michael L., "Microphone array processing for robust speech recognition.", Carnegie Mellon University, 2003.
- [9] Deng Jia, Dong Wei et al., "Imagenet: A large-scale hierarchical image database", 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 2009, 248-255.
- [10] WU, Bichen, et al. "Visual transformers: Token-based image representation and processing for computer vision.", arXiv preprint arXiv:2006.03677, 2020.
- [11] HE, Kaiming, et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. p. 770-778.
- [12] Dimauro Giovanni, Girardi Francesco, "Italian Parkinson's Voice and Speech", *IEEE DataPort*, 10.21227/aw6b-tg17, 2019.
- [13] Dimauro Giovanni, Di Nicola Vincenzo et al., "Assessment of Speech Intelligibility in Parkinson's Disease Using a Speech-To-Text System", *IEEE Access*, 5, 22199-22208, 2017.
- [14] Dimauro Giovanni, Caivano Danilo et al., "VoxTester, software for digital evaluation of speech changes in Parkinson disease", 2016 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Benevento, Italy, 2016, 1-6.
- [15] Prior Fred, Virmani Tuhin et al., "Voice Samples for Patients with Parkinson's Disease and Healthy Controls", figshare, 10.6084/m9.figshare.23849127.v1, 2023.