Automatic Parkinson's disease detection from speech: Layer selection vs adaptation of foundation models

Tilak Purohit*†, Barbara Ruvolo *, Juan Rafael Orozco-Arroyave ‡, Mathew Magimai.-Doss *
*Idiap Research Institute, Martigny, Switzerland
†EPFL, École polytechnique fédérale de Lausanne, Switzerland
‡GITA Lab, Universidad de Antioquia, Colombia
{tilak.purohit, mathew.magimaidoss}@ idiap.ch

Abstract—**In this work, we investigate Speech Foundation Models (SFMs) for Parkinson's Disease (PD) detection. We explore two main approaches: (1) using SFMs as frozen feature extractors and, (2) fine-tuning/adapting SFMs for PD detection. We propose a cross-validation-based layer selection methodology to identify the layer effective for PD detection. Additionally, we compare the performance of the layer selection scheme with full fine-tuning and, parameter-efficient fine-tuning (PEFT) using Low-Rank Adaptation (LoRA). Our results show that layer selection and LoRA-based fine-tuning can perform on par with full fine-tuning, providing a more parameter-efficient alternative. The highest accuracy was achieved by fine-tuning Whisper using LoRA.**

Index Terms—**Parkinson's Disease, Speech for health, Foundation Models, PEFT, LoRA, Fine-tuning, PC-GITA.**

I. Introduction

Parkinson's disease (PD), a neurodegenerative disorder caused by the progressive loss of dopaminergic neurons [1], often results in speech impairments such as reduced voice quality, monotonicity, and difficulty in articulation [2], [3]. Speech analysis offers a non-invasive, cost-effective approach for automatic PD detection, motivating the development of systems to reduce the time and effort required for clinical assessments of PD-related speech disorders, dysarthria [4], [5].

Traditional methods for dysarthria detection have utilized acoustic features like jitter, shimmer, formants, glottal features and Mel Frequency Cepstral Coefficients (MFCCs) [6]–[8]. With the success of Deep Learning (DL) in various fields [9] there has been a surge in research focused on using DL for automatic pathological speech detection [10], [11]. These approaches employ various speech representations and architectures. For instance, [12]–[14] utilized Convolutional Neural Networks (CNNs), while [15]–[17] applied Long Short-Term Memory (LSTM) networks. Rios-Urrego et al. [18] used a convolutional recurrent network, consisting of a 1D-CNN followed by an LSTM, to classify PD versus healthy controls (HC). The authors in [19], [20] utilized latent features such as i-vectors [21] and x-vectors [22], originally developed for speaker verification and identification, to differentiate PD from HC. Wodzinski et al. [23] and Karaman et al. [24] identified the limited availability of pathological speech data for training deep neural networks and thus employed a transfer learning approach for PD detection. Both studies utilized the ResNet architecture [25], pretrained on the ImageNet corpus [26]. Wodzinski used the pretrained features for classification, while Karaman fine-tuned the network for PD detection. Although the pretrained network was meant for image classification both the studies demonstrated the effectiveness of transfer learning for the PD detection task.

The scarcity of diverse, publicly available pathological speech data remains a challenge. However, the success of **S**peech **F**oundation **M**odels (SFMs) [27] in various speech-related downstream tasks has led to increased interest in leveraging transfer learning for pathological speech analysis. Recent studies [28]–[31] highlights the effectiveness of SFMs, particularly wav2vec2.0-base [32], in encoding different speech pathologies.

When using SFMs, there are two primary approaches: freezing them to serve as feature extractors or fine-tuning/adapting them for downstream tasks, as illustrated in Figure 1 (a) and Figure 1 (b), respectively. Studies have demonstrated that when using SFMs as feature extractors, each layer captures distinct speech-related information [33], [34], making layer selection a potentially advantageous approach for the task. Furthermore, full fine-tuning/adaptation of SFMs for pathological speech is still underexplored. In this work, we focus on the PD detection task utilising SFMs and in the scope of this work we: (a) propose a methodology for layer selection in SFMs, (b) investigate the effectiveness of adaptation on SFMs, and (c) investigate LoRA-based adaptation approach for PD detection, utilizing parameter-efficient fine-tuning (PEFT) via the Low-Rank Adaptation (LoRA) method [35]. While LoRA has been explored in various speech processing tasks [36]–[38], its application to pathological speech detection remains unexplored.

Specifically, we aim to assess whether adapting large SFMs with massive parameter spaces is feasible for pathological speech in a data-constrained scenario, or if utilizing features from a particular layer selected through our proposed methodology is sufficient. We also explore if PEFT approach of strategically updating a small subset of parameters within the SFM (using LoRA) is more efficient than full fine-tuning for the PD detection task.

This work was partially funded by (a) the SNSF through the Bridge Discovery project EMIL: Emotion in the loop - a step towards a comprehensive closed-loop deep brain stimulation in Parkinson's disease (grant no. $40B2-0\_194794$), and (b) CODI at UdeA, grant # PI2023-58010.

The rest of the paper is organised as following, Section II introduces the methods investigated in this study. Section III outlines the dataset used and the experimental setup, Section IV presents the experimental results and the analysis, finally Section V concludes the paper.

II. Methods Investigated

A. Cross validation based layer selection

We propose a simple approach, (i) Extract the representations (of speech utterances) from different layers of the SFMs (ii) train an auxiliary classifier for the task-on hand and evaluate only using the cross-validation set (iii) The classifier's accuracy on the cross-validation set serves as an indicator of the layer's effectiveness in learning task-related properties. (iv) the best performing layer on the cross-validation set is selected to evaluate the test set.

B. Fine-tuning/Adaptation

Fine-tuning/adaptation involves updating all model parameters, including those of the foundation model (upstream model) and the classifier block (downstream model), to tailor the network for the target task, as shown in Fig 1(b).

C. LoRA

LoRA is a PEFT technique, proposed to efficiently adapt large FMs to specific domains or downstream tasks. Consider $W_0 \in \mathbb{R}^{d\times k}$ to be the pre-trained weight matrix, LoRA replace model update with low-rank matrix decomposition as follows, $W_0 + \Delta W = W_0 + BA$, where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and rank $r \ll min(d,k)$. During training W_0 is kept frozen while A and B are trainable parameters as shown in Figure 1(c) . This strategy significantly reducing the number of trainable parameters.

III. Experimental Setup

A. Dataset and Protocol

We consider PC-GITA corpus [39] for the PD detection task, the corpus consists of 100 participants, divided equally between 50 Parkinson's disease (PD) patients and 50 healthy controls (HC), all native Spanish speakers from Colombia. The two groups are balanced in terms of age, gender, and education level. Each participant contributed 10 sentences and a phonetically balanced text, providing an average of 55.5 seconds of speech data per participant. Originally the speech data was recorded at a sampling frequency of 44.1 kHz, we downsampled the recordings to 16 kHz for our study.

For our investigation, we resort to a stratified 10-fold speaker-independent cross-validation evaluation. At each fold, 80%, 10%, and 10% of the data is used for training, cross-validation, and testing, respectively. Following the previous work, we report the performance of our classification systems in terms of accuracy, F1-score, sensitivity (correct classification rate for PD), and specificity (correct classification rate for HC). The final performance is the mean and standard deviation of classification metric values obtained across 10 folds of the test-set. It is worth noting that with this protocol, we acquire around 1 hour and 15 minutes of speech data per fold for training.

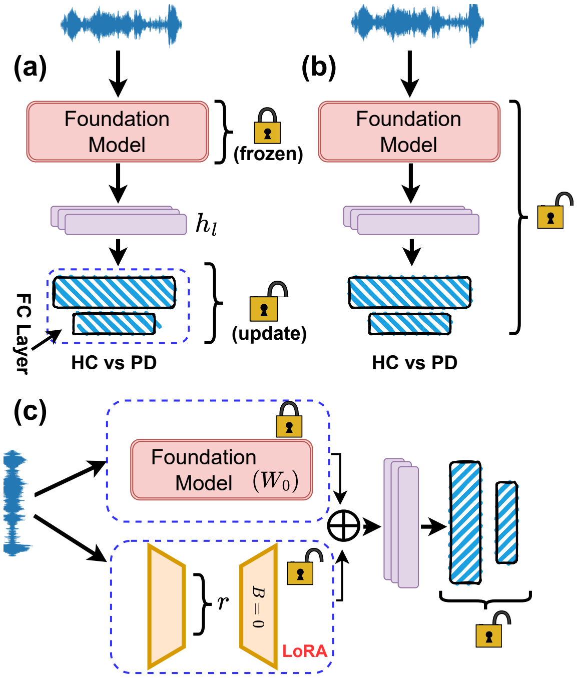


Fig. 1: *Figure depicting different training methodologies: (a) Linear probing, (b) Fine-Tuning, and (c) Fine-Tuning using LoRA ; 'h_l' and 'FC Layer' refer to frame-level embeddings from layer 'l' and the fully connected layer, respectively; HC= healthy control and PD= Parkinson's disease*

B. System description and Configurations

(a) Handcrafted features: For Baseline we utilize knowledge-based feature representations provided with openSMILE toolkit [40]. We utilize ComParE$_{LLD\times F}$, which consists of 6,373 static turn-level features derived from computing functionals (statistics) over $(65+65)$ LLD contours. Additionally, we conducted experiments with eGeMAPS$_{LLD\times F}$, which includes 88 static turn-level features obtained from functionals computed over 23 LLD contours.

We use support vector machine (SVM) as a classifier for the handcrafted feature based pipeline. SVM performance was optimized by doing grid search for hyperparameters such as the kernel, kernel width (γ) and soft margin constant (C) using the cross-validation set.

(b)Speech Foundation Models: In this work, we examine three SFMs for the PD detection task. We carefully selected these systems to enable a detailed analysis of various training approaches, such as self-supervised and weakly supervised methods. Also, the impact of monolingual versus multilingual pretraining, this is relevant since the PC-GITA corpus consists of Spanish recordings.

Wav2vec2.0-base [32] hereafter referred to as W2V2, utilizes a self-supervised learning approach, combining contrastive learning with masking techniques. It comprises 12 transformer encoder layers, 768-dimensional hidden states, and 8 attention

TABLE I: *Comparison of different feature representations for PD vs. HC classification results on test-set, averaged over 10-folds on PC-GITA.(·) indicates the standard deviation. "Param." indicate network's trainable parameters for respective systems*

Feature representation	Param.	Dim.	Accuracy ↑	F1-score ↑	Sensitivity ↑	Specificity ↑
G-1: Handcrafted Features						
ComParE$_{LLD\times F}$	–	6373	77.60(7.1)	77.62(7.1)	76.90(10.2)	79.16(12.3)
eGeMAPS$_{LLD\times F}$	–	88	76.44(6.3)	75.98(6.4)	75.45(11.8)	77.45(13.3)
G-2: Cross-validation layer selection for SFMs (selected layer)						
W2V2 (L 10)	$197K$	768	83.54(5.6)	83.73(5.3)	84.72(9.8)	82.36(13.4)
XLSR (L 16)	$263K$	1024	83.72(8.3)	84.12(8.1)	86.12(10.6)	81.27(13.2)
WHISPER (L 12)	$197K$	768	81.09(8.6)	81.93(8.0)	86.00(10.7)	76.18(12.1)
G-2.2: Combining decisions from all layers of SFMs						
W2V2	$197K$	768	84.27(6.5)	84.54(6.2)	85.82(9.4)	82.73(12.5)
XLSR	$263K$	1024	83.91(7.1)	84.45(6.7)	86.73(8.6)	81.09(12.9)
WHISPER	$197K$	768	75.73(9.6)	74.27(9.9)	70.73(12.7)	80.73(15.5)
G-3: Fine-tuning SFMs						
W2V2-FT	$90.4M$	768	80.90(8.3)	80.09(9.7)	79.27(15.6)	82.54(13.1)
XLSR-FT	$311M$	1024	84.72(7.8)	85.51(7.2)	89.45(8.6)	80.00(12.8)
WHISPER-FT	$87.2M$	768	83.53(8.3)	83.91(8.3)	86.06(11.5)	81.01(14.4)
G-4: Fine-tuning SFMs with LoRA adapters (rank)						
W2V2 (R 4)	$862K$	768	83.09(7.2)	83.06(7.4)	83.81(12.2)	82.36(13.1)
XLSR (R 4)	$2M$	1024	83.09(7.5)	83.30(7.3)	84.90(12.4)	81.27(15.4)
WHISPER (R 16)	$2.9M$	768	85.00(9.6)	85.34(8.9)	86.36(10.1)	83.63(15.2)

heads, amounting to 95 million parameters. Pretraining of network was conducted on 960 hours of English audio using the LibriSpeech corpus.

XLSR [41] is a multilingual variant of Wav2vec2.0 with 24 transformer encoder layers, 1024-dimensional hidden states, and 16 attention heads. The model is pretrained on 53 languages and consist of 315M parameters.

Whisper-small [42] henceforth denoted as whisper, the network was pretrained in a wealky-supervised fashion. The model comprises 12 encoder and 12 decoder blocks, each with 12 attention heads and a 768-dimensional hidden state, totaling 244 million parameters. Whisper is multilingual, and was trained on approximately 680,000 hours of weakly-supervised speech data sourced from the internet.

All the SFMs were retrieved from HuggingFace. The frame level representation derived from the SFMs were mean pooled and then fed to the classifier head consisting of one hidden layer with 256 nodes and output layer of 2 nodes corresponding to the number of classes, 2 in this case (HC and PD). The output layer had softmax activation, while the hidden layer had ReLU activation. The networks were trained using cross-entropy loss with Adam optimizer. The batch size was set to 4, with gradient accumulation configured at 8, and the seed value was set to 1337. When probing the network for the downstream task (Figure 1(a)), the learning rate was set to 1×10^{-4}. For fine-tuning (Figure 1(b)), the learning rate was adjusted to 1×10^{-5}. It is worth emphasizing during fine-tuning, the CNN-based encoder blocks were kept frozen for all SFMs, while only the transformer encoder blocks were fine-tuned. In the case of whisper, the 12 encoder layers were used to generate speech representations, while the decoder layers were excluded. Lastly, for fine-tuning with LoRA (Figure 1(c)), we performed experiments with rank (r) value across the set {4, 8, 16} for each SFMs the best performing results are reported here.

IV. Results and Analysis

A. System performance

Figure 2 present the accuracy trends for the three SFMs based on the layer-wise analysis performed on the cross-validation set on 10 folds. The results reveals that the best-performing layer for W2V2 is Layer 10, which aligns with findings from previous studies [28], [29]. For XLSR, Layer 16 shows the highest performance, while Whisper's optimal layer is Layer 12. Notably, XLSR exhibits minimal fluctuation in performance across the mid layers, with the last layer performing the worst. Table I G-2 presents the test set outcomes for the layers selected using the cross-validation set. W2V2 and XLSR achieves higher accuracy, while XLSR and Whisper show better sensitivity scores. Comparing layer selection to fine-tuning, shown in Table I G-3, we observe a slight accuracy improvement for XLSR and whisper, but a decrease for W2V2. Additionally, the sensitivity score for XLSR increases after fine-tuning, while it decreases for W2V2 and remains unchanged for Whisper. When comparing layer selection results with LoRA adaptation

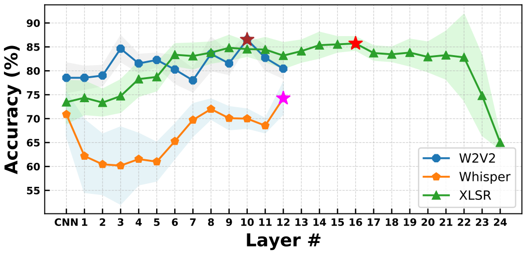


Fig. 2: -●- *on curves depicts mean of classification accuracy over 10 folds on validation-set at every layer. Best Accuracy (on validation-set data):* -★- *W2V2: 86.53%,* -★- *Whisper: 74.27%, and* -★- *XLSR: 85.72% from layer 10, 12 and 16 respectively*

in Table I G-4, we observe a 4% absolute gain in whisper's accuracy. Furthermore, the results indicate that SFM-based classification whether using layer selection or fine-tuning, consistently outperforms the handcrafted features, reported in Table I G-1. Notably, the best-performing accuracy of 85.00% obtained with whisper finetuned using LoRA with rank value 16 (in Table I G4) surpasses what has been previously reported in the literature using a similar train test protocol for PC-GITA- 83% [14] and 82.6% [29]. When compared to previous work [43] which investigated PD classification on PC-GITA using only fine-tuning, our study presents a comparative analysis, demonstrating that selecting the appropriate layer and using PEFT-based fine-tuning can achieve performance comparable to full fine-tuning.

B. Analysis

Layer selection analysis: To better analyse our cross validation layer selection scheme, we computed the layer-wise accuracy for the test-set similarly to what we did for cross-validation data. Figure 3 showcases the accuracy trend for test-set. We see a some common layer-wise accuracy trend for both test and cross-validation data. For W2V2 on test data layer-3 yield 85.36% accuracy, whereas using cross-validation layer selection layer-10 is selected, yielding an accuracy of 83.54% on test data, which is not a large difference. If observed for cross-validation set (in Figure 2) next best pick would have been layer-3. For whisper the trend remains the same with lower layer not performing well, and the last layer (layer-12) being the best pick for both validation and test set. This seems valid as whisper is trained for speech recognition (SR) task and last layer being task specific for SR could pick up on the cues of atypical PD speech. For XLSR via cross validation layer selection layer-16 is picked which has test set accuracy of 83.72% whereas layer-15 on test data yields 87.06% accuracy, this might be an anomaly observed from the peak at layer-15 in Figure 3. Otherwise, for XLSR the accuracy trend remains the same for the test and the cross-validation data, that is the middle layer yield better results, and for the last layer a drastic drop in accuracy is observed. This analysis showcases that cross-validation layer selection scheme generalise well to the test set. We also analyze the impact of combining decisions from all layers using a majority voting strategy. Table I (G2.2) presents the results, where no significant performance gain is observed, with a noticeable drop in Whisper's performance. These findings, along with the trends shown in Figure 2 and 3, suggest that selectively choosing layers for combining decisions may be more advantageous.

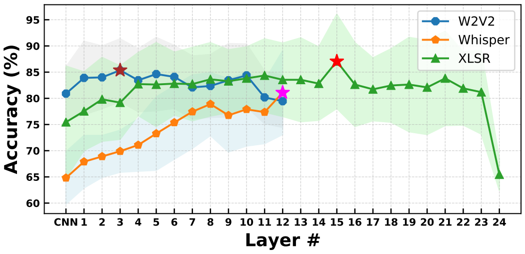
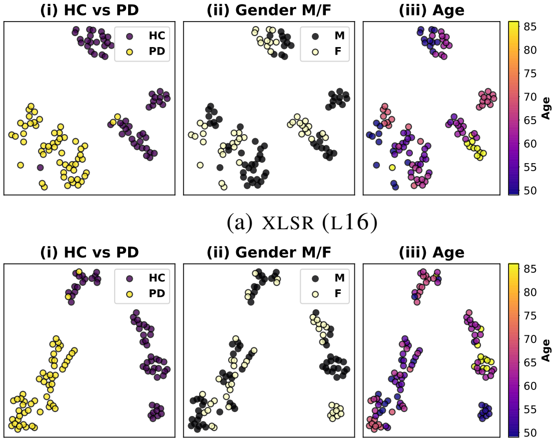


Fig. 3: -●- *on curves depicts mean of classification accuracy over 10 folds on test-set at every layer. Best Accuracy (on test set data):* -★- *W2V2: 85.36%,* -★- *Whisper: 81.09%, and* -★- *XLSR: 87.06% from layer 3, 12 and 15 respectively*



(a) XLSR (L16)

(b) WHISPER (R 16)

Fig. 4: *t-SNE plot of the last layer embedding space from selected systems, using 110 utterances from 10 speakers (5 HC and 5 PD) in the test set. Data points are color-coded to represent: (i) HC vs PD, (ii) Gender (M and F), and (iii) Age.*

Embedding space analysis: We generate embeddings from the test set data of one fold, consisting of 110 utterances from 10 speakers (5 HC and 5 PD), with 3 males and 2 females in each group. For embedding generation, we select two systems: XLSR (L 16) with 83.72% accuracy, and Whisper (R 16) fine-tuned using LoRA, achieving 85% accuracy. To visualize the embedding space, we use t-SNE plots, as shown in Figure 4. For XLSR (L 16) and Whisper (R 16), we observe distinct clusters for PD and HC. When these clusters are color-coded by gender and age, we notice distinct and systematic clustering, particularly within the HC set. This indicates that these networks might also be capturing speaker identity information in the process of PD detection.

V. Conclusion

In this work, we explore Speech Foundation Models (SFMs) for the task of PD detection. We introduce a cross-validation-based layer selection methodology and compare its effectiveness to full fine-tuning or adaptation of the SFMs. Additionally, we for the first time employ LoRA-based fine-tuning for PD detection. Our results show that the layer selection approach achieves accuracy on par with, and sometimes equal to, full fine-tuning, offering a more parameter cost-efficient alternative. LoRA adaptation for whisper outperforms layer selection, possibly because whisper is pre-trained for speech recognition, and fine-tuning it with LoRA enables it to detect atypical speech, caused due to articulation difficulties in PD patients. This naturally leads to future work, where we plan to explore SFMs fine-tuned for ASR tasks in the context of PD detection.

References

[1] O. Hornykiewicz, "Biochemical aspects of parkinson's disease," *Neurology*, 1998.

[2] K. K. Baker, L. O. Ramig, E. S. Luschei, and M. E. Smith, "Thyroarytenoid muscle activity associated with hypophonia in parkinson disease and aging," *Neurology*, 1998.

[3] J. Rusz, R. Cmejla, T. Tykalova, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, and E. Ruzicka, "Imprecise vowel articulation as a potential early marker of parkinson's disease: effect of speaking task," *The Journal of the Acoustical Society of America*, 2013.

[4] T. Virmani, M. Lotia, A. Glover, L. Pillai, A. S. Kemp, A. Iyer, P. Farmer, S. Syed, L. J. Larson-Prior, and F. W. Prior, "Feasibility of telemedicine research visits in people with parkinson's disease residing in medically underserved areas," *Journal of Clinical and Translational Science*, 2022.

[5] Q. C. Ngo *et al.*, "Computerized analysis of speech and voice for parkinson's disease: A systematic review," *Computer Methods and Programs in Biomedicine*, 2022.

[6] T. Bocklet, S. Steidl, E. Nöth, and S. Skodda, "Automatic evaluation of parkinson's speech-acoustic, prosodic and voice related cues." in *Interspeech*, 2013.

[7] N. N. Prabhakera and P. Alku, "Dysarthric speech classification using glottal features computed from non-words, words and sentences," in *Interspeech*, 2018.

[8] S. Hawi, J. Alhozami, R. AlQahtani, D. AlSafran, M. Alqarni, and L. El Sahmarany, "Automatic parkinson's disease detection based on the combination of long-term acoustic features and mel frequency cepstral coefficients (MFCC)," *Biomedical Signal Processing and Control*, 2022.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, 2015.

[10] R. Gupta, T. Chaspari, J. Kim, N. Kumar, D. Bone, and S. Narayanan, "Pathological speech processing: State-of-the-art, current challenges, and future directions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[11] N. Cummins, A. Baird, and B. W. Schuller, "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning," *Methods*, 2018.

[12] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, and E. Nöth, "Convolutional neural network to model articulation impairments in patients with parkinson's disease." in *Interspeech*, 2017.

[13] I. Kodrasi, "Temporal envelope and fine structure cues for dysarthric speech detection using CNNs," *IEEE Signal Processing Letters*, 2021.

[14] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Automatic dysarthric speech detection exploiting pairwise distance-based convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[15] A. A. Joshy and R. Rajan, "Automated dysarthria severity classification using deep learning frameworks," in *European Signal Processing Conference (EUSIPCO)*, 2021.

[16] T. Bhattacharjee, J. Mallela, Y. Belur, A. Nalini, R. Yadav, P. Reddy, D. Gope, and P. K. Ghosh, "Source and vocal tract cues for speech-based classification of patients with parkinson's disease and healthy subjects." in *Interspeech*, 2021.

[17] S. Bhati, L. M. Velazquez, J. Villalba, and N. Dehak, "LSTM siamese network for parkinson's disease detection from speech," in *IEEE global conference on signal and information processing*, 2019.

[18] C. D. Rios-Urrego, S. A. Moreno-Acevedo, E. Nöth, and J. R. Orozco-Arroyave, "End-to-end parkinson's disease detection using a deep convolutional recurrent network," in *International Conference on Text, Speech, and Dialogue*, 2022.

[19] N. Garcia, J. R. Orozco-Arroyave, D. Luis Fernando, N. Dehak, and E. Nöth, "Evaluation of the neurological state of people with parkinson's disease using i-vectors." in *Interspeech*, 2017.

[20] L. Moro-Velazquez, J. Villalba, and N. Dehak, "Using x-vectors to automatically detect parkinson's disease from speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[21] N. Dehak, P. J. Kenny, R. OPT, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[22] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[23] M. Wodzinski, A. Skalski, D. Hemmerling, J. R. Orozco-Arroyave, and E. Nöth, "Deep learning approach to parkinson's disease detection using voice recordings and convolutional neural network dedicated to image classification," in *Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, 2019.

[24] O. Karaman, H. Çakın, A. Alhudhaif, and K. Polat, "Robust automated parkinson disease detection based on voice signals with transfer learning," *Expert Systems with Applications*, 2021.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, 2015.

[27] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe *et al.*, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[28] D. Wagner, I. Baumann, F. Braun, S. P. Bayerl, E. Nöth, K. Riedhammer, and T. Bocklet, "Multi-class detection of pathological speech with latent features: How does it perform on unseen data?" in *Interspeech*, 2023.

[29] M. Amiri and I. Kodrasi, "Adversarial robustness analysis in automatic pathological speech detection approaches," in *Interspeech*, 2024.

[30] D. A. Wiepert, R. L. Utianski, J. R. Duffy, J. L. Stricker, L. R. Barnard, D. T. Jones, and H. Botha, "Speech foundation models in healthcare: Effect of layer selection on pathological speech feature prediction," in *Interspeech*, 2024.

[31] D. Escobar-Grisales, C. D. Ríos-Urrego, I. Baumann, K. Riedhammer, E. Noeth, T. Bocklet, A. M. Garcia, and J. R. Orozco-Arroyave, "It's time to take action: Acoustic modeling of motor verbs to detect parkinson's disease," in *Interspeech*, 2024.

[32] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, 2020.

[33] A. Pasad, B. Shi, and K. Livescu, "Comparative layer-wise analysis of self-supervised speech models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[34] S. A. Chowdhury, N. Durrani, and A. Ali, "What do end-to-end speech models learn about speaker, language and channel information? a layer-wise and neuron-level analysis," *Computer Speech & Language*, 2024.

[35] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[36] T. Feng and S. Narayanan, "Peft-ser: On the use of parameter efficient transfer learning approaches for speech emotion recognition using pre-trained speech models," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2023.

[37] W. Liu, Y. Qin, Z. Peng, and T. Lee, "Sparsely shared Lora on whisper for child speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[38] Z. Song *et al.*, "Lora-whisper: Parameter-efficient and extensible multilingual ASR," in *Interspeech*, 2024.

[39] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, "New spanish speech corpus database for the analysis of people suffering from parkinson's disease." in *Lrec*, 2014.

[40] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. of ACM Multimedia*, 2010.

[41] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *Interspeech*, 2021.

[42] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*, 2023.

[43] M. La Quatra *et al.*, "Exploiting foundation models and speech enhancement for parkinson's disease detection from speech in real-world operative conditions," in *Interspeech*, 2024.