On Detection of Depression in Parkinson's Disease Patients' Speech: Handcrafted Features vs. Speech Foundation Models

Tilak Purohit[1,2], Barbara Ruvolo[1], Juan Rafael Orozco-Arroyave[3], and Mathew Magimai.-Doss[1]

[1] Idiap Research Institute, Martigny, Switzerland
[2] EPFL, École polytechnique fédérale de Lausanne, Switzerland
[3] GITA Lab, Universidad de Antioquia, Colombia
{tilak.purohit, mathew.magimaidoss}@idiap.ch

Abstract. This study investigates speech-based depression detection in individuals with Parkinson's disease (PD), comparing two feature representation approaches: interpretable handcrafted acoustic features and non-interpretable representations derived from Speech Foundation Models (SFMs). The study utilizes the DAIC-WOZ corpus for typical depression (non-PD speech) and the PD-Depression corpus for depression in atypical speech (PD speech). We first examine the viability of handcrafted features and then analyse how acoustic descriptors differ across these conditions. We then evaluate SFM-based representations on the PD-Depression dataset using a cross-validation-based layer selection methodology. Results suggest that handcrafted features yield better classification performance for depression in PD speech compared to SFM embeddings. Analysis shows that while typical depression is marked by pitch instability and reduced vocal quality, speech from PD patients with depression exhibits broader spectral variability, likely due to motor impairments associated with hypokinetic dysarthria.

Keywords: Depression detection · Parkinson's disease · Interpretable features · Speech Foundation Models.

1 Introduction

Depressive disorder is a common mental health condition that affects an estimated 5% of the global adult population, according to the World Health Organization [41]. It is associated with a range of emotional, cognitive, and behavioral symptoms [38], such as persistent low mood, anxiety, and slowed motor activity, and in some cases may lead to suicidal thoughts [1]. Diagnosis typically depends on clinical interviews, which can vary in consistency and may contribute to heterogeneous assessments [6, 24, 4]. This highlights the relevance of developing reliable automated approaches to support the screening of depression across its diverse manifestations.

In the speech community, numerous studies have explored the use of acoustic features for detecting depression, demonstrating the potential of vocal indicators [28, 15, 27, 7, 39, 43]. These studies typically involve statistical analysis of features such as fundamental frequency (F0), intensity, and various spectral properties, suggesting their relevance as possible markers for depression screening. In recent years, artificial neural network (ANN)-based methods have become increasingly popular for this task, showing promising results in capturing speech patterns linked to depressive states [42, 19, 26, 9]. While these methods often outperform traditional handcrafted approaches in terms of accuracy, they generally lack interpretability and are constrained by the availability of sufficient labeled data for training.

Although the field has seen substantial advances, accurately detecting depression from speech remains difficult due to the diverse manifestations of depression. One area that has received comparatively less attention is the investigation of depression when it co-occurs with neurological disorders such as Parkinson's disease or Alzheimer's disease [25]. Aarsland et al.[2] emphasize that while Parkinson's disease (PD) is mainly recognized for its motor impairments, it also includes a wide range of non-motor symptoms, depression being among the most common, affecting about one-third of individuals with PD. Notably, depression in PD is often persistent and may even manifest during the prodromal phase[10].

The coexistence of depression with neurological disorders such as PD adds complexity to the diagnostic process and often results in underdiagnosis and inadequate treatment [21]. Despite growing interest in speech-based mental health assessment, the detection of depression from atypical speech remains an underexplored area. To date, only a few studies have specifically addressed this challenge. Ozkanca et al. [30] were among the first to investigate depression screening in PD patients using brief 10-second phoneme recordings. Their approach relied on handcrafted acoustic features combined with standard machine learning classifiers, demonstrating a promising relationship between vocal patterns and depression in PD. However, the study was limited in scope to isolated phoneme-level speech. In contrast, Pérez-Toro et al. [32–34] focused on longer speech samples, analyzing monologues from individuals with PD and Alzheimer's disease. They applied transfer learning strategies by first modeling affective states using ForestNet [35] in the valence-arousal space and then fine-tuning the models for depression detection. While their work demonstrates the potential of using emotion modeling as an intermediate task, it still leaves open questions regarding direct depression classification from atypical, disorder-specific speech. Together, these studies highlight a notable gap in the literature: the limited investigation into depression detection from neurologically atypical speech signals, where both motor and affective symptoms may alter vocal expression.

This study builds upon the work of Ruvolo et al. [36], who examined whether interpretable handcrafted acoustic features commonly used in automatic speech-based depression detection could be effectively applied to detect depression in PD patients. Their analysis also compared acoustic descriptors from non-PD and PD speech to identify factors that might hinder accurate speech depression

classification. For completeness, we present their findings and expand on their work by incorporating representations from state-of-the-art Speech Foundation Models (SFMs) to further explore the challenges of modeling depression in PD speech using such representations.

Using handcrafted features, we analyzed and compared the acoustic characteristics of individuals with comorbid Parkinson's disease and depression to those with depression alone, using two distinct corpora. In addition, we introduced a feature filtering method based on the Point Biserial Correlation Coefficient (PBCC) to identify the most informative features for depression detection, aiming to enhance classification performance. Finally, for clarity, we categorize the acoustic features into three main groups based on the type of information they represent: vocal source-related features, which influence pitch, loudness, and voice quality; vocal tract-related features, which reflect modifications in sound as it travels through the vocal tract, such as formants; and global-related features, which capture overall processes influencing speech production, like energy.

SFM-derived representations were used exclusively for detecting depression in individuals with Parkinson's disease. To assess the viability of these representations, we adopted the cross-validation-based layer selection approach proposed by Purohit et al. [31], originally developed for Parkinson's disease detection task.

The remainder of the paper is structured as follows: Section 2 describes the datasets used in the study. Section 3 outlines the methods explored, while Section 4 details the feature representations and experimental setup. Section 5 presents the experimental results, followed by a discussion and analysis in Section 6. Finally, Section 7 concludes the paper.

2 Dataset

(a) Depression in Parkinson's disease (PD-D) [32]: The dataset comprises speech recordings from 60 Colombian Spanish speakers, including 25 Depressive Parkinson's Disease (D-PD) patients and 35 Non-Depressive Parkinson's Disease (ND-PD) patients. Each participant was asked to deliver a monologue describing their daily routine. Following the recordings, a neurologist assessed their neurological condition using the Movement Disorders Society – Unified Parkinson's Disease Rating Scale (MDS–UPDRS) [16], which is the standard tool for evaluating the neurological state of PD patients. The initial section of the MDS-UPDRS includes an item that evaluates depression based on daily routine activities, assigning scores from 0 to 4. Participants scoring above zero were categorized as D-PD, while those with a score of zero were identified as ND-PD. The average length of the monologues is 84±34 seconds for D-PD and 80±37 seconds for ND-PD, resulting in a total dataset duration of approximately 4892 seconds. Speaker 52 was excluded from the analysis due to recording issues.

(b) Distress analysis interview corpus - Wizard of Oz (DAIC-WOZ) [17]: The dataset includes audio-visual interviews from 189 participants who were evaluated for psychological distress, amounting to 17 hours of audio data. Each participant was assigned a self-assessed depression score based on the Patient

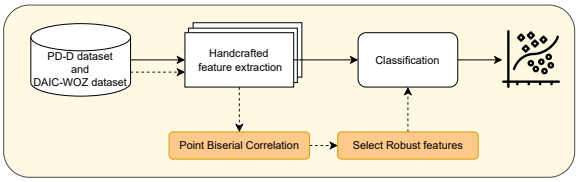

Fig. 1. Proposed methodologies representation: conventional approach (solid arrows) and PBCC-based approach (dashed arrows).

Health Questionnaire (PHQ-8) method [23]. For our experiments, we extracted only the participants' speech recordings using the time labels provided in the dataset. Sessions 318, 321, 341, and 362 were excluded from the training set due to time-labelling errors.

Table 1 summarizes the two datasets.

Table 1. Distribution of utterances used in the study, corresponding to each label.

Database	Content	Depressed patients	Not-Depressed patients	Total
DAIC-WOZ	English	42	100	142
PD-D	Spanish	24	35	59

3 Methodology

This study employs two distinct approaches, which are outlined below.

3.1 Feature selection for handcrafted acoustic descriptors

As a baseline approach, handcrafted features were initially extracted from the input audio signal and used as input representations for the classifier module to produce confidence scores (see solid arrows in Fig. 1). This method provides a useful point of comparison with more sophisticated architectures.

Building upon the handcrafted features approach, we introduce a feature selection step to refine the input representation. After extracting features from the raw audio signal, we apply the Point Biserial Correlation Coefficient (PBCC) to identify the most informative subset of features (see dashed arrows in Fig. 1). PBCC measures the strength of the linear association between a binary target variable (D/ND) and continuous-valued features, enabling the selection of features that are most relevant for classification.

The PBCC is defined as:

$$r_{pb} = \frac{M_D - M_{ND}}{s_n} \sqrt{\frac{n_D n_{ND}}{n^2}}$$

where M_D and M_{ND} denote the mean feature values for the Depressive (D) and Non-Depressive (ND) classes, respectively; n_D and n_{ND} are the number of samples in each class; n is the total number of samples; and s_n is the standard deviation of the feature values across all samples.

To begin with, a range of correlation thresholds is defined. For each threshold value, the PBCC is calculated between every feature and the target labels in the training set. Features that exceed the given threshold are retained, as they demonstrate a stronger linear association with the class labels and are considered more discriminative.

Next, a Gradient Boosting (GB) classifier is employed to evaluate the predictive performance of the selected feature subsets. Thresholds ranging from 0.06 to 0.6, with increments of 0.2, are tested to determine which subset yields the highest classification accuracy.

3.2 Cross validation based layer selection for SFMs

The proposed architecture consists of two main components: an upstream SFM encoder and a downstream classification module. The upstream encoder comprises a stack of N transformer layers, each incorporating self-attention mechanisms. The downstream module includes an average pooling layer followed by a fully connected multilayer perceptron (MLP).

To determine which encoder layer produces the most task-relevant representations, we adopt a simple yet effective procedure: (i) Representations are extracted from each layer of the SFM encoder using the input speech utterances. (ii) The downstream network is trained using these representations to perform the required task, and its performance is evaluated exclusively on the cross-validation set. (iii) The accuracy on the cross-validation set serves as a proxy for the effectiveness of the representations in capturing task-specific information. (iv) The layer that achieves the highest cross-validation accuracy is selected for final evaluation on the test set.

We now detail the training procedure for the proposed model.

Let X represent the input speech signal and $c \in \{D, ND\}$ the corresponding class label. The objective is to estimate the posterior probability $P(c|X)$

$$\begin{aligned} H_L &= \mathrm{ENC}_L(X; \theta_{e_L}), \\ P(c|X) &= \mathrm{softmax}(\mathrm{MLP}(\mathrm{Pool}(H_L); \theta_m)) \ \forall c \end{aligned} \tag{1}$$

Here, $H_L \in \mathbb{R}^{T \times D}$ denotes the sequence of embeddings from the L-th layer of the encoder, where T is the sequence length and D the embedding dimension. $\mathrm{ENC}_L(\cdot)$ is the encoder function for layer $L \in \{1, \ldots, N\}$ with parameters θ_{e_L}. The pooling function $\mathrm{Pool}(\cdot)$ aggregates the sequence of embeddings into a fixed-length representation, which is then passed through the MLP, parameterized by θ_m, followed by a softmax to produce the class probabilities.

The predicted label $\hat{c}$ is assigned by selecting the class with the highest posterior probability:

$$\hat{c} = \arg\max_{c} P(c|X)$$

Model training is guided by minimizing the cross-entropy loss $\mathcal{L}$ between the predicted distribution $P(c|X)$ and the ground truth label c:

$$\mathcal{L} = -\log P(c|X)$$

The encoder parameters θ_{e_L} are initialized using weights from a pretrained model and remain frozen during training. In contrast, the downstream parameters θ_m are randomly initialized and optimized during training.

4 Feature description and training protocol

In this section, we provide a brief overview of the handcrafted features and SFMs used in this study, along with the training protocol applied for each setup.

4.1 Handcrafted features

This study utilizes three widely recognized sets of knowledge-driven handcrafted features: eGeMAPS [13], ComParE [37], and DisVoice [29]. The eGeMAPS set comprises 25 Low-Level Descriptors (LLDs) that capture essential acoustic properties, including frequency, energy, amplitude, and spectral characteristics. These descriptors are further processed using a set of statistical functionals, yielding a total of 88 features. ComParE provides a significantly larger feature set, comprising 6373 descriptors that include delta coefficients at the frame level and statistical functionals at the utterance level, all extracted via the openSMILE toolkit [12]. DisVoice features, on the other hand, combine static representations related to phonation, articulation, and prosody into a unified feature vector. Further details and extraction procedures are available in [29].

For the classification task, we employed three machine learning algorithms: Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting (GB). Hyperparameters for each model were optimized using grid search in conjunction with 6-fold cross-validation, ensuring reliable and robust model selection.

To maintain consistency with prior work [32], we employed the Leave-One-Speaker-Out (LOSO) cross-validation protocol for the PD-D corpus. Specifically, for evaluating the k-th speaker, the model was trained on data from the remaining $k-1$ speakers. For the DIAC dataset, evaluations were conducted on the development set, as the official test set was reserved for the AVEC 2016 challenge [40].

4.2 SFM based features

In this work, we investigate two Speech Foundation Models (SFMs) for the task of PD-D detection. These models were deliberately chosen to facilitate a comparative analysis of the impact of monolingual versus multilingual pertaining, a

relevant factor given that the PD-D corpus comprises Spanish-language recordings.

Wav2vec2.0-base [3], hereafter referred to as W2V2, adopts a self-supervised learning framework that integrates contrastive learning with masked prediction. The model architecture consists of 12 transformer encoder layers with 768-dimensional hidden states and 8 attention heads, totaling approximately 95 million parameters. It was pretrained on 960 hours of English speech from the LibriSpeech corpus.

XLSR [5] is a multilingual extension of Wav2vec2.0, featuring 24 transformer encoder layers, 1024-dimensional hidden states, and 16 attention heads, resulting in a model size of 315 million parameters. It was pretrained on speech data spanning 53 languages.

The Speech Foundation Models (SFMs) were obtained from HuggingFace. The downstream head composed of a single hidden layer with 256 nodes and an output layer of 2 nodes, corresponding to the two target classes (Healthy Control [HC] and Parkinson's Disease [PD]). The hidden layer employed ReLU activation, while the output layer used a softmax activation to produce class probabilities. Model training was performed using the Adam optimizer and cross-entropy loss. A batch size of 4 was used, with gradient accumulation set to 8 and the seed value was set to 1337. During probing for the downstream classification task, a learning rate of 5×10^{-4} was used for W2V2, and 1×10^{-4} for XLSR.

To evaluate the PD-D corpus using the SFM layer selection approach, we employed a 5-fold cross-validation strategy. Each fold consisted of speaker-independent training, validation, and test sets, comprising 35, 12, and 12 samples respectively.

5 System performance

This section reports the results and examines the classification performance associated with each feature category.

5.1 Hancrafted features

Table 2 summarizes the results obtained using the best-performing classifier (Gradient Boosting, GB) for each feature set across both corpora considered in this study. In the DAIC-WOZ dataset, our proposed system substantially outperforms the baseline results reported in the AVEC 2016 challenge, which were based on the eGeMAPS feature set with an SVM classifier, achieving an improvement of approximately 51% in F1-score. While the conventional use of eGeMAPS yields an F1-score of 0.74, performance slightly drops to 0.69 following feature selection, likely due to the limited dimensionality of the feature set. Conversely, both ComPARE and *DisVoice* show marked performance gains after applying feature selection.

For the PD-D corpus, our methods surpass the overall F1-score of 0.70 reported by Pérez-Toro et al. [32], who employed Valence and Arousal representations for classification. Among the conventional approaches, *DisVoice* features

Table 2. Classifiers' performance over the two datasets. *Dims* denotes the feature dimension; *Thr.* signifies the threshold set for feature selection; *D* and *ND* denote depressed and not-depressed patients, respectively; *O* is the unweighted average of *D* and *ND*.

			F1-score			Precision		Recall	
Features	**Dims**	**Thr.**	**O**	**D**	**ND**	**D**	**ND**	**D**	**ND**
DAIC-Woz									
Valstar et al. [40]	88		0.49	0.41	0.58	0.26	0.94	0.88	0.42
Conventional approach									
eGeMAPS	88		**0.74**	0.62	0.87	0.90	0.78	0.47	0.97
ComPARE	6373		0.47	0.24	0.70	0.45	0.59	0.16	0.86
DisVoice	620		0.55	0.33	0.77	0.50	0.69	0.25	0.87
Feature-selection approach									
eGeMAPS	39	0.18	0.69	0.54	0.84	0.67	0.72	0.33	0.91
ComPARE	2756	0.18	**0.72**	0.65	0.80	0.62	0.78	0.33	0.87
DisVoice	184	0.20	0.65	0.47	0.83	0.80	0.73	0.33	0.96
PD-D									
Perez-Toro et al. [32]			0.68	-	-	-	-	-	-
Conventional approach									
eGeMAPS	88		0.54	0.40	0.69	0.53	0.61	0.32	0.79
ComPARE	6373		0.43	0.30	0.56	0.33	0.53	0.28	0.59
DisVoice	620		**0.74**	0.69	0.78	0.71	0.77	0.68	0.79
Feature-selection approach									
eGeMAPS	7	0.26	0.65	0.57	0.72	0.62	0.68	0.52	0.76
ComPARE	186	0.3	**0.78**	0.75	0.81	0.73	0.82	0.76	0.79
DisVoice	16	0.32	0.77	0.73	0.81	0.75	0.80	0.72	0.82

initially exhibit the highest performance with an F1-score of 0.74. However, applying Point-Biserial Correlation Coefficient (PBCC)-based feature selection leads to significant improvements across all feature sets. In particular, ComPARE improves from 0.43 to 0.78 while using a reduced subset of just 186 features.

Overall, these findings demonstrate that PBCC-based feature selection effectively enhances classifier performance, especially in identifying depressed (D) patients. The only exception is the slight decline observed for eGeMAPS in the DAIC-WOZ dataset. These results highlight the benefit of filtering out redundant or non-informative features, enabling the development of more compact, interpretable, and discriminative models.

5.2 SFM based features

Figure 2 illustrates the accuracy trends of the two SFMs based on layer-wise analysis conducted over 5-fold cross-validation. The results indicate that Layer 6 yields the highest performance for W2V2, while Layer 16 is optimal for XLSR. Table 3 presents the test-set results on the PD-D corpus using the layers selected through the cross-validation-based layer selection strategy. The findings clearly

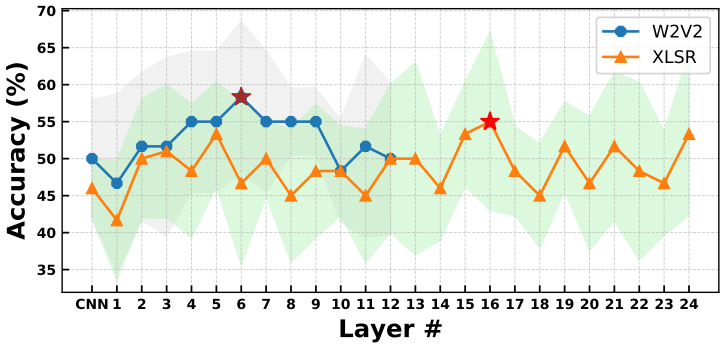

Fig. 2. -●- *on curves depicts mean of classification accuracy over 5 folds on validation-set at every layer. Best Accuracy (on validation-set data):* -★- *W2V2: 58.33%, and* -★- *XLSR: 55.00% from layer 6, and 16 respectively*

indicate that the SFM-derived representations alone do not yield satisfactory performance for this task. Notably, the XLSR model pretrained on multiple languages including Spanish achieves an unweighted average F1-score of 40%, whereas the W2V2 model, pretrained solely on English, attains a higher score of 45%.

Table 3. SFM performance for the selected layer on test-set, *Dims* denotes the feature dimension; *D* and *ND* denote depressed and not-depressed patients, respectively; *O* is the unweighted average of *D* and *ND*.

		F1-score			Precision		Recall	
Features	**Dim.**	**O**	**D**	**ND**	**D**	**ND**	**D**	**ND**
				PD-D				
W2v2 (L 6)	768	0.45	0.19	0.70	0.43	0.58	0.12	0.88
XLSR (L 16)	1024	0.40	0.12	0.67	0.29	0.56	0.08	0.85

6 Result analysis and discussion

This section offers a detailed analysis and interpretation of the results reported in the previous section.

6.1 Hancrafted features

Building on the classification results from handcrafted features, we carried out a feature importance analysis by identifying the top 10 features according to their normalized importance scores, as determined by the best-performing classifier

Table 4. Feature ranking of GB trained on ComPARE for both DAIC-WOZ (left) and PD-D (right), using PBCC feature selection approach.

DAIC-WOZ			PD-D		
Index	**LLD name**	**Group**	**Index**	**LLD name**	**Group**
3861	logHNR	Source	142	Length L1	Global
4038	Jitter DDP	Source	88	RMS Energy	Global
5126	Py Sharpness	Source	1	Length L1	Global
1493	Spectral Harmonicity	Source	64	RMS Energy	Global
6077	Spectral Variance	Global	6067	Spectral Entropy	Vocal tract
6174	MFCC	Vocal tract	1245	Spectral Flux	Global
2365	audSpec	Vocal tract	1476	Py Sharpness	Source
4132	F0	Source	2925	Spectral RollOff 90.0	Vocal tract
4131	F0	Source	2991	Spectral Centroid	Global
1373	Spectral Skeweness	Vocal tract	3106	Spectral Kurtosis	Global

from the feature selection pipeline on both datasets. To facilitate interpretation, we organized these acoustic descriptors into three distinct categories based on the type of information they represent. Following the classification framework proposed by Eyben et al.[11], we divided Low-Level Descriptors (LLDs) into those related to the vocal source and those linked to the vocal tract. Additionally, we introduced a third category encompassing features that capture global properties of the speech signal, integrating information from both source and tract. Table 4 lists the selected features in descending order of importance, as assigned by the Gradient Boosting (GB) model. For reference, the descriptor names are provided, with the 'Index' column indicating the position of the feature in the `openSMILE` feature list, and the 'Group' column showing the assigned category.

The feature rankings in Table 4 reveal a distinct pattern in how the classifier identifies depression in speech from individuals with and without Parkinson's disease. In the DAIC-WOZ corpus, which involves classifying depressive versus healthy control subjects, the classifier places the greatest emphasis on source-related features—accounting for 6 out of the top 10. This is followed by a smaller number of vocal tract features (3 out of 10) and just one global low-level descriptor (LLD). The prominence of source-related features suggests that vocal characteristics tied to the stability and quality of vocal fold vibrations such as: jitter, harmonic-to-noise ratio, and pitch variability are particularly informative for depression detection in a neurologically healthy population. These results align with clinical findings, such as those by Hollien [20], who reported pitch alterations in depressed individuals, and Darby [8], who noted diminished pitch range and vocal intensity among depressed patients. In contrast, the feature rankings for the PD-D dataset indicate a broader spread of informative features. Notably, the presence of several spectral features such as: spectral entropy, spectral centroid, spectral roll-off, spectral flux, and spectral kurtosis—reflects the vocal instability often observed in individuals with Parkinson's disease [18]. The observed reductions in spectral entropy and spectral centroid among depressed patients further highlight the nuanced interplay between emotional states and vocal quality [22]. Additionally, the feature Length L1, which captures aspects of voice quality and consistency, is significantly impacted, pointing to irregular speech patterns likely caused by motor impairments. Overall, this distribution

underscores the added complexity in detecting depression in PD patients due to the interplay between emotional and motor-related vocal alterations.

6.2 SFM based features

While handcrafted acoustic features enable interpretability and facilitate detailed analysis, such transparency is not readily achievable with representations derived from SFMs. Interestingly, contrary to earlier findings where SFM-based features have outperformed handcrafted ones for tasks like depression detection [42] and Parkinson's disease classification [14, 31], this performance advantage diminishes in scenarios where depression coexists with Parkinson's disease. One possible explanation lies in the limitations of SFMs, which are typically pretrained on large-scale general speech corpora and may thus lack sensitivity to the subtle interplay between affective cues and motor deficits characteristic of this comorbid condition. The performances obtained further support this observation, for instance, when we analyse the mean sensitivity (correct classification rate for PD-D) and specificity (correct classification rate for PD-ND) scores across five folds, we observe a high specificity of 82% and a notably low sensitivity of 24% for W2V2 (L 6). This imbalance hints at the model's bias toward the majority class and its difficulty in capturing information related to depression within the pathological speech condition. Overfitting to the training data may also contribute to this limitation. To address these issues, future work will explore data augmentation and balancing strategies aimed at improving model robustness and sensitivity to underrepresented classes.

7 Conclusion

This study investigates automatic depression detection in individuals with Parkinson's disease (PD) using the PD Depression (PD-D) corpus. We explore two distinct feature types for this work: (1) interpretable, knowledge-based handcrafted acoustic representations, and (2) non-interpretable features derived from Speech Foundation Models (SFMs). For the handcrafted feature analysis, we not only evaluate their performance on the PD-D corpus but also study and compare acoustic descriptor patterns with those observed in typical depression-affected speech using the DAIC-WOZ corpus. The evaluation of SFM-derived features is focused exclusively on the PD-D dataset to assess their effectiveness in detecting depression within atypical speech associated with PD.

The classification results indicate that handcrafted feature-based approaches, such as those using the ComPARE feature set, can achieve an F1-score of 43% for depression detection in individuals with Parkinson's disease. Notably, performance improves substantially with the F1-score rising to 78% when redundant features are removed using a Point Biserial Correlation-based feature selection strategy. These findings highlight the critical role of feature selection in enhancing model robustness and discriminative power. In contrast, the performance of SFM-derived features remains subpar, even when representations are

extracted from the optimal encoder layer identified via cross-validation layer selection methodology. The results indicate that these models tend to exhibit a bias toward the majority class, indicating limited effectiveness in capturing depression-related cues within the atypical speech patterns of individuals with Parkinson's disease.
The analysis of feature rankings (using handcrafted features) highlights distinct acoustic profiles of depression in non-PD versus PD speech. In the DAIC-WOZ dataset, the classifier places strong emphasis on source-related features, underscoring their critical role in detecting depressive states an insight consistent with prior studies showing that vocal fluctuations, such as pitch variability, are reliable markers of depression. Whereas, speech from individuals with Parkinson's disease exhibits a wider prominence of spectral features, reflecting the intricate speech alterations associated with the motor symptoms characteristic of the disorder.

Our findings suggest that the presence of Parkinson's disease complicates the automatic classification of depression, likely due to overlapping and interacting acoustic symptoms. The study demonstrates that depression manifests differently in the speech of individuals with Parkinson's compared to those without, emphasizing the need for tailored approaches in speech-based mental health assessment. Our results demonstrate that handcrafted features tend to yield better PD-D detection than SFM-derived representations. To better leverage the potential of SFMs, future work will focus on exploring data augmentation strategies aimed at addressing both the limited data availability and class imbalance inherent in the PD-D dataset.

One limitation of this study is the reliance on a low-resource and imbalanced dataset, where depression labeling is based solely on a non-zero score for the "Depressed mood" item in the MDS–UPDRS scale. We believe, a more robust labeling criterion would be necessary to enhance the reliability of the ground truth. Additionally, the dataset does not include information on medication status or speech-related subscores, both of which could provide valuable context for experimental design and interpretation of results. Addressing these limitations would require a separate dedicated data collection in the future.

Acknowledgment

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.

References

1. Global health estimates: Life expectancy and leading causes of death and disability. https://www.who.int/health-topics/mental-health, last accessed 15/01/2025

2. Aarsland, D., Påhlhagen, S., Ballard, C., Ehrt, U., Svenningsson, P.: Depression in parkinson disease—epidemiology, mechanisms and management. Nature Reviews Neurology (2012)
3. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems (2020)
4. Burdick, B., Holmes, C., Waln, R.: Recognition of suicide signs by physicians in different areas of specialization. Journal Of Medical Education (1983)
5. Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M.: Unsupervised cross-lingual representation learning for speech recognition. Interspeech (2021)
6. Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., Quatieri, T.: A review of depression and suicide risk assessment using speech analysis. Speech Communication (2015)
7. Cummins, N., et al.: An investigation of depressed speech detection: Features and normalization. In: Proc. of INTERSPEECH (2011)
8. Darby, J., Simmons, N., Berger, P.: Speech and voice parameters of depression: A pilot study. Journal Of Communication Disorders (1984)
9. Dubagunta, S., Vlasenko, B., Doss, M.: Learning voice source related information for depression detection. In: Proc. of ICASSP (2019)
10. Dušek, P., et al.: Relations of non-motor symptoms and dopamine transporter binding in rem sleep behavior disorder. Nature Scientific reports (2019)
11. Eyben, F., Eyben, F.: Acoustic features and modelling (2016)
12. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings Of The 18th ACM International Conference On Multimedia (2010)
13. Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., et al.: The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. IEEE transactions on affective computing (2015)
14. Favaro, A., Tsai, Y.T., Butala, A., Thebaud, T., Villalba, J., Dehak, N., Moro-Velázquez, L.: Interpretable speech features vs. dnn embeddings: What to use in the automatic assessment of parkinson's disease in multi-lingual scenarios. Computers in Biology and Medicine (2023)
15. France, D., et al.: Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Transactions On Biomedical Engineering (2000)
16. Goetz, C., Tilley, B., Shaftman, S., Stebbins, G., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M., Dodel, R., Others: Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results. Movement Disorders: Official Journal Of The Movement Disorder Society (2008)
17. Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Others: The distress analysis interview corpus of human and computer interviews. In: LREC (2014)
18. Hauptman, Y., Aloni-Lavi, R., Lapidot, I., Gurevich, T., Manor, Y., Naor, S., Diamant, N., Opher, I.: Identifying distinctive acoustic and spectral features in parkinson's disease. In: Proc. of INTERSPEECH (2019)
19. He, L., Cao, C.: Automated depression analysis using convolutional neural networks from speech. J. Biomed. Inform. (2018)
20. Hollien, H.: Vocal indicators of psychological stress. Annals Of The New York Academy Of Sciences (1980)

21. Hussain, M., et al.: Similarities between depression and neurodegenerative diseases: pathophysiology, challenges in diagnosis and treatment options. Cureus (2020)
22. Hussenbocus, A., et al.: Statistical differences in speech acoustics of major depressed and non-depressed adolescents. In: Proc. of ICSPCS (2015)
23. Kroenke, K., Strine, T.W., Spitzer, R.L., Williams, J.B., Berry, J.T., Mokdad, A.H.: The phq-8 as a measure of current depression in the general population. Journal of affective disorders (2009)
24. Landau, M.: Acoustical properties of speech as indicators of depression and suicidal risk. Vanderbilt Undergraduate Research Journal (2008)
25. Lee, H., Lyketsos, C.: Depression in alzheimer's disease: heterogeneity and related issues. Biological Psychiatry (2003)
26. Ma, X., et al.: Depaudionet: An efficient deep model for audio based depression classification. In: Proc. of AVEC, On ACM MM (2016)
27. Moore II, E., et al.: Critical analysis of the impact of glottal features in the classification of clinical depression in speech. IEEE Transactions On Biomedical Engineering (2007)
28. Nilsonne, A., et al.: Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression. The Journal of the Acoustical Society of America (1988)
29. Orozco-Arroyave, J.R., Vásquez-Correa, J.C., Vargas-Bonilla, J.F., Arora, R., Dehak, N., Nidadavolu, P.S., Christensen, H., Rudzicz, F., Yancheva, M., Chinaei, H., et al.: Neurospeech: An open-source software for parkinson's speech analysis. Digital Signal Processing (2018)
30. Ozkanca, Y., Göksu Öztürk, M., Ekmekci, M., Atkins, D., Demiroglu, C., Hosseini Ghomi, R.: Depression screening from voice samples of patients affected by parkinson's disease. Digital Biomarkers (2019)
31. Purohit, T., Ruvolo, B., Orozco-Arroyave, J.R., Magimai.-Doss, M.: Automatic parkinson's disease detection from speech: Layer selection vs adaptation of foundation models. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2025)
32. Pérez-Toro, P., Vásquez-Correa, J., Bocklet, T., Nöth, E., Orozco-Arroyave, J.: User state modeling based on the arousal-valence plane: Applications in customer satisfaction and health-care. IEEE Transactions On Affective Computing (2021)
33. Pérez-Toro, P., et al.: Depression assessment in people with parkinson's disease: The combination of acoustic features and natural language processing. Speech Communication (2022)
34. Pérez-Toro, P., et al.: Transferring quantified emotion knowledge for the detection of depression in alzheimer's disease using forestnets. In: Proc. of ICASSP (2023)
35. Rodríguez-Salas, D., et al.: Mapping ensembles of trees to sparse, interpretable multilayer perceptron networks. SN Computer Science (2020)
36. Ruvolo, B., Purohit, T., Vlasenko, B., Orozco-Arroyave, J., Doss, M.: Exploring the complexity of parkinson's patient speech for depression detection task: A qualitative analysis. In: Proceedings of Workshop on Speech Pathology Analysis and DEtection (SPADE 2025) (2025)
37. Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., et al.: The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In: Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France (2013)

38. Scibelli, F., Roffo, G., Tayarani, M., Bartoli, L., Mattia, G., Esposito, A., Vinciarelli, A.: Depression speaks: Automatic discrimination between depressed and non-depressed speakers based on nonverbal speech features. In: Proc. of ICASSP (2018)
39. Stasak, B., et al.: An investigation of emotional speech in depression classification. In: Proc. of INTERSPEECH (2016)
40. Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., Pantic, M.: Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In: AVEC (2016)
41. World Health Organization: Depression. https://www.who.int/news-room/fact-sheets/detail/depression, last accessed 19/10/2023
42. Wu, W., Zhang, C., Woodland, P.C.: Self-supervised representations in speech-based depression detection. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2023)
43. Zahid, L., et al.: Detection of speech impairments in parkinson disease using hand-crafted feature-based model on spanish speech corpus. In: Proc. of CCIS (2020)