Federated Interpretable Speech Biomarkers for Parkinson's Detection: Evaluating Explanation Stability

Najmeh Mohajeri[1,2] [0009−0003−0428−3730] and Ute Bauer-Wersing[1] [0009−0004−6094−4545]

[1] Frankfurt University of Applied Sciences, Frankfurt am Main, Germany
ubauer@fra-uas.de
[2] Metropolia University of Applied Sciences, Helsinki, Finland
najmeh.mohajeri@metropolia.fi

Abstract. Parkinson's disease (PD) produces early, measurable changes in speech, making acoustic analysis an attractive non-invasive route to screening. The datasets needed to train reliable models are scattered across clinics in different countries, and because voice recordings are identifiable and subject to privacy regulation they cannot be pooled freely. We present a proof-of-concept framework that couples federated learning (FL) with clinically interpretable speech biomarkers, so that three simulated institutions jointly train a PD classifier without exchanging any raw recordings. The feature set is anchored on the vowel space area (VSA), an articulatory marker that contracts under PD-related vowel centralization, together with standard phonatory descriptors. A logistic model is trained by federated averaging; because its inputs are standardized clinical measures, its coefficients double as a global explanation of which biomarkers drive each decision. On synthetic non-IID data that emulates cross-site differences in recording conditions and vocal-tract characteristics, and across 30 independent seeds, the federated model reaches a pooled AUC of 0.831 ± 0.010. This is statistically equivalent to a data-pooling centralized model (two one-sided tests, ± 0.02 bound, $p \approx 2 \times 10^{-24}$) and, in a paired comparison, exceeds the 0.822 mean of siloed local models ($\Delta = +0.009$, paired t, $p < 10^{-6}$). VSA is the top-ranked feature in 29 of 30 seeds and all coefficient signs match established physiology. Our contribution is a transparent articulation-grounded pipeline whose explanation is exact and stable, together with a modest-but-consistent federation benefit under heterogeneity. The framework is a reproducible template for later validation on real multilingual corpora; it makes no clinical performance claim.

Keywords: Parkinson's disease · Federated learning · Speech biomarkers · Vowel space area · Explainable AI · Data-minimizing machine learning

1 Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disorder, and up to four in five patients develop phonatory and articulatory symptoms, often before a formal diagnosis [7]. These changes, termed hypokinetic dysarthria, make recorded speech a promising, inexpensive substrate for early screening, and machine-learning models on acoustic features increasingly separate patients from healthy controls [7]. Impaired vowel articulation in particular is a sensitive early indicator, detectable even before a clinician perceives dysarthria [9].

Progress is limited less by algorithms than by data. Reliable models need large, diverse corpora, yet labelled recordings sit in single institutions, countries and languages; because voice is biometric and identifiable, sharing it across borders raises legal and ethical barriers that centralized training cannot overcome [3]. Federated learning (FL) circumvents this: the model is sent to each site, trained locally, and only its parameters are aggregated [1]. Federated models can approach centrally trained quality without any data leaving the hospitals [2], and the paradigm has recently been applied to PD speech across languages [4,5].

Two gaps motivate this work. First, existing federated PD-speech systems emphasize accuracy, often through opaque deep representations, whereas clinical uptake also demands decisions that are explainable and checkable against known physiology [12]. Second, the articulatory markers clinicians trust most, such as the vowel space area, are rarely the backbone of federated pipelines.

We therefore contribute: (1) a compact FL pipeline built on interpretable articulatory and phonatory biomarkers centred on the vowel space area; (2) an intrinsically explainable federated classifier whose standardized coefficients give a global, clinically verifiable account of its decisions, which we show recovers the expected sign of every biomarker and the primacy of VSA robustly across 30 seeds; and (3) a controllable synthetic non-IID benchmark with open code that, under paired statistical testing, isolates a small but significant federation advantage over local-only training while confirming equivalence to a data-pooling upper bound. The empirical contribution of this study lies in demonstrating the stability of the model's intrinsic explanations and the benefits of federated learning under non-IID data distributions,two properties that are not guaranteed by the convergence theory of the Federated Averaging (FedAvg) algorithm. This work is a methodological proof of concept based on synthetic data and does not make any claims regarding clinical performance.

2 Related Work

Speech and Vowel Biomarkers of PD. Acoustic assessment of PD spans phonatory measures such as jitter, shimmer and harmonics-to-noise ratio, and articulatory measures derived from vowel formants [7]. Among the latter, the vowel space area (VSA) and the related vowel articulation index and formant centralization ratio quantify how far the corner vowels /a/, /i/ and /u/ are pushed toward the centre of the formant plane [8,10]. PD contracts this space through articulatory

undershoot, an effect observable in early and even prodromal stages [9], and it can now be extracted automatically in a language-independent manner [11]. Such markers suit a federated setting because they are low-dimensional, physiologically grounded and directly interpretable.

Federated Learning for Medical and PD Speech Data. FL has become a practical paradigm for privacy-sensitive medical data, with server–client aggregation as the dominant design and safeguards such as secure aggregation and differential privacy layered on top [2,3]. For PD, federated models trained on German, Spanish and Czech speech outperformed local models and approached centralized performance without data sharing [4]; related studies give federated proofs of concept and optimize aggregation for the strong non-IID heterogeneity of multilingual corpora [5,6]. Our work is complementary: rather than optimizing a deep model, we ask how far a deliberately transparent, biomarker-driven federated model can go while staying explainable.

Explainable AI for PD Speech. Post-hoc attribution, most commonly Shapley additive explanations, has been used to expose which acoustic features drive PD predictions [12,13]. Whereas these methods provide post hoc explanations for otherwise opaque models, we employ an intrinsically interpretable classifier whose parameters themselves serve as the explanation. This approach avoids the fidelity gap inherent to surrogate models while remaining compatible with Shapley analysis for independent validation and cross-checking.

3 Method

3.1 Interpretable Speech Biomarkers

Each speaker is represented by five standardized biomarkers. The central articulatory feature is the vowel space area, computed from the first two formants (F_1, F_2) of the corner vowels as the area of the triangle they span,

$$\mathrm{VSA} = \tfrac{1}{2}\,\big|\, F_{1a}(F_{2i} - F_{2u}) + F_{1i}(F_{2u} - F_{2a}) + F_{1u}(F_{2a} - F_{2i}) \,\big|. \qquad (1)$$

A smaller VSA indicates vowel centralization and thus greater articulatory impairment [8,10]. Four phonatory descriptors complete the set: jitter and shimmer (cycle-to-cycle frequency and amplitude perturbation), harmonics-to-noise ratio (HNR), and the standard deviation of F_0 (F_0 SD), which captures reduced pitch range (monopitch).

3.2 Federated Learning Setup

We consider $K = 3$ institutions, each holding a private set of n_k labelled speakers. Every site trains a logistic classifier $\sigma(w^\top x + b)$ on its local data for E epochs and returns its full parameter vector $\theta_k = (w_k, b_k)$, comprising both the weight

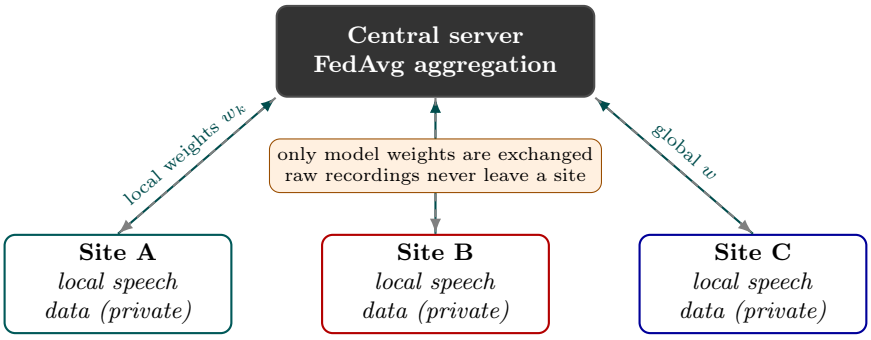

Fig. 1. Federated setup: each institution trains on its own private speech data and shares only model weights w_k; the central server returns the aggregated global model. Raw recordings never leave a site.

vector and the bias term, to a central server. The server forms the global model by federated averaging weighted by local sample count,

$$\theta^{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} \theta_k^{t+1}, \qquad n = \sum_{k=1}^{K} n_k, \tag{2}$$

so that the weights and the bias are aggregated together. The updated global parameters are broadcast back and the cycle repeats for T rounds [1]. No raw recordings or features ever leave a site; only parameter vectors are exchanged (Fig. 1). Features are standardized with global per-feature means and variances, which in a real deployment are obtained from aggregated per-site sums and counts via secure aggregation rather than by pooling data.

3.3 Interpretability

Because all inputs are standardized, each coefficient's magnitude is a comparable measure of that biomarker's global importance, and its sign shows whether increasing the biomarker pushes the decision toward PD or non-PD. The explanation is thus exact for the deployed model, needs no surrogate, and can be read against clinical expectations, while remaining compatible with Shapley analysis for per-patient explanations [13].

4 Experimental Setup

To study interpretability and behaviour reproducibly, we generate synthetic data with a known ground truth; this avoids premature clinical claims, lets us vary the degree of cross-site heterogeneity, and sidesteps corpus licensing. For each speaker the label is drawn first, then formants are generated so that PD speakers show larger articulatory undershoot, pulling the corner vowels toward the centre

and shrinking the VSA, with speaker-level noise that makes the classes overlap. Non-IID heterogeneity is introduced by scaling the whole vowel space per site, emulating language- and vocal-tract-dependent differences: the three sites differ in vowel-space scale, PD prevalence, and sample size ($n = 200, 130, 200$, with PD prevalence $0.45, 0.50, 0.55$ respectively); phonatory features are drawn from overlapping PD/healthy distributions.

Each of the $K = 3$ sites trains a logistic classifier by full-batch gradient descent with learning rate 0.30 and L_2 regularization 10^{-3}, for $E = 15$ local epochs per round over $T = 20$ federated rounds; the centralized and local-only baselines are trained for $T \cdot E = 300$ epochs under the same optimizer for a fair comparison. Each site is split 70/30 into train and test partitions. The class-conditional phonatory distributions (PD vs. healthy) are jitter $\mathcal{N}(0.50, 0.16)$ vs. $\mathcal{N}(0.42, 0.16)$; shimmer $\mathcal{N}(0.30, 0.09)$ vs. $\mathcal{N}(0.26, 0.09)$; HNR $\mathcal{N}(19.5, 3.3)$ vs. $\mathcal{N}(21.0, 3.3)$; F_0 SD $\mathcal{N}(19.0, 6.5)$ vs. $\mathcal{N}(23.0, 6.5)$. On a pooled test set we compare three regimes: a *centralized* model on the union of the training data (a privacy-violating upper bound), the mean of three isolated *local-only* models, and the *federated* model, reporting the area under the ROC curve (AUC) and accuracy. To quantify variability we repeat the entire pipeline over 30 seeds and report means with 95% confidence intervals. We test federated-vs-centralized equivalence with two one-sided tests (TOST, ± 0.02 AUC bound) and federated-vs-local superiority with a paired t-test across seeds.

5 Results and Discussion

The synthetic biomarker behaves as intended: averaged over speakers and seeds, the PD vowel triangle is markedly smaller than the healthy one (156k versus 235k Hz^2 for VSA, a $\sim$34% contraction), reproducing the vowel-centralization effect reported clinically (Fig. 2).

Across 30 seeds the federated model reaches a pooled-test AUC of 0.831 ± 0.010 (Fig. 3a). Two results follow. First, the federated model is statistically equivalent to the data-pooling centralized model: the paired per-seed difference is essentially zero and falls well inside a ± 0.02 equivalence bound (TOST, $p \approx 2 \times 10^{-24}$, Fig. 3b). We treat this as a sanity check rather than a contribution: the classifier is convex, so federated averaging converges to the same optimum a centralized fit would reach, and the equivalence is theoretically expected. Second, and less trivially, federation yields a small but statistically significant improvement over siloed local training. Because both regimes are evaluated on the same per-seed data, we compare them pairwise: the mean advantage is $+0.009$ AUC (paired t, $p < 10^{-6}$), positive in 26 of 30 seeds (Fig. 3b). The marginal confidence intervals of the two regimes overlap, but the paired difference is consistent and significant; the effect reflects the non-IID setting, where a model fitted to one site's vowel-space scale transfers imperfectly to the others while federation exposes the shared model to all three regimes without moving data. Per-site federated AUC varies across the three sites in line with their differing distributions rather than any failure of aggregation. The absolute gain is modest and we do

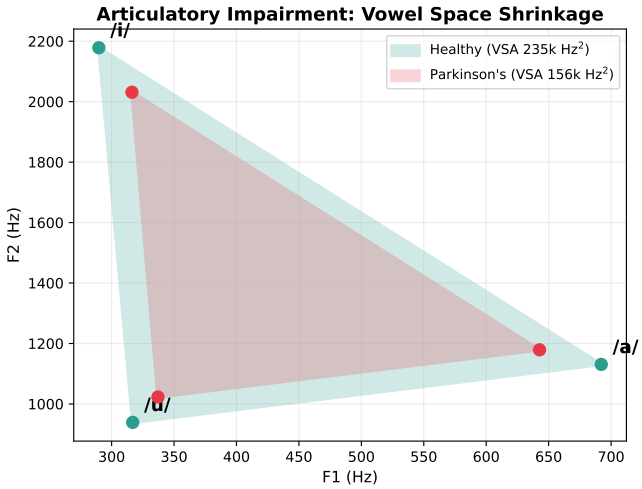

Fig. 2. Mean vowel triangles in the F_1–F_2 plane, averaged over 30 seeds. Parkinsonian speech centralizes the corner vowels /a/, /i/ and /u/, contracting the vowel space area.

not overstate it: on this synthetic benchmark the practical value of federation over local training is real but small, and its magnitude on real heterogeneous corpora remains an open question. The pooled accuracy of the federated model is 0.755.

The model is explainable by construction, and the explanation is stable. Its standardized coefficients (Fig. 4) rank VSA as the most influential biomarker in 29 of 30 seeds, with a negative sign—a larger vowel space pushes the decision toward the healthy class, exactly as the articulatory literature predicts [8,9]. All five signs are recovered consistently: elevated jitter and shimmer push toward PD, while higher HNR and F_0 variability push toward the healthy class, each with 95% intervals that exclude zero. We are precise about the scope of this stability. What is robust is the *direction* of every biomarker and the *primacy* of VSA; the finer rank order among the four phonatory descriptors is not stable across seeds and should not be over-interpreted. This qualified but auditable alignment between an intrinsically interpretable federated model and established physiology is obtained without any post-hoc approximation, and can be cross-checked with Shapley analysis for per-patient explanations [13].

These findings echo prior federated PD-speech work, in which federation recovers most of the benefit of pooling while preserving privacy [4], and add an intrinsically interpretable, articulation-grounded formulation. Limitations follow from the design. The results are synthetic and carry no clinical claim; the classifier is intentionally simple; and the federation advantage, while significant, is small on this benchmark. Crucially, our privacy claim is limited to data minimization: raw recordings and raw features never leave a site. Sharing logistic coefficients and global standardization statistics can leak information about the

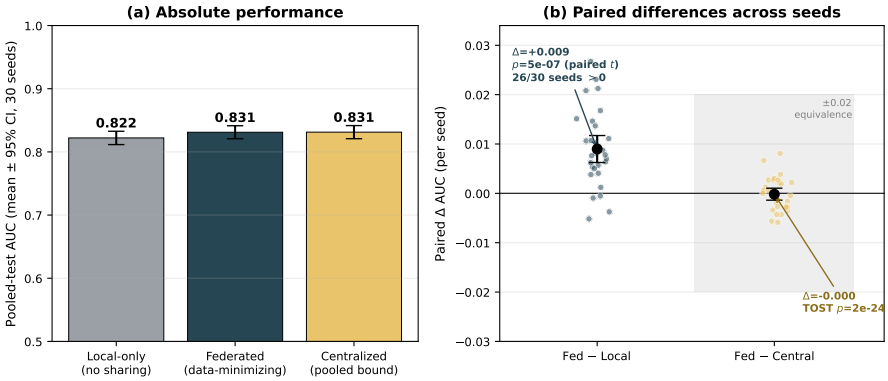

Fig. 3. (a) Pooled-test AUC of the three regimes over 30 seeds (mean ± 95% CI). (b) Paired per-seed differences. Federated training exceeds local-only by +0.009 AUC (paired t, $p < 10^{-6}$; 26/30 seeds positive) and is statistically equivalent to the centralized upper bound, falling inside a ±0.02 equivalence band (TOST, $p \approx 2 \times 10^{-24}$). The paired view shows the local-vs-federated gap is significant despite overlapping marginal intervals.

training data, and we make no differential-privacy or secure-aggregation guarantee, these mechanisms are assumed, not realized. We therefore describe the system as *data-minimizing* and *privacy-motivated* rather than privacy-preserving. Next steps are: validation on real multilingual corpora ; a leakage analysis and formal differential-privacy guarantees; more expressive yet interpretable models; and a Shapley cross-check of the coefficient explanations.

6 Conclusion

We presented a proof-of-concept uniting federated learning, interpretable speech biomarkers and intrinsic explainability for cross-institutional PD detection. On controlled non-IID data the federated model matched a centralized upper bound—an expected consequence of convexity that we treat as a sanity check—and delivered a small but statistically significant improvement over isolated local models while sharing no raw data. Its transparent parameters recovered the clinically expected primacy and direction of vowel-space contraction robustly across seeds, offering an auditable starting point for validation on real clinical corpora.

Acknowledgments. This work used only synthetic data. The authors thank colleagues for helpful discussions.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

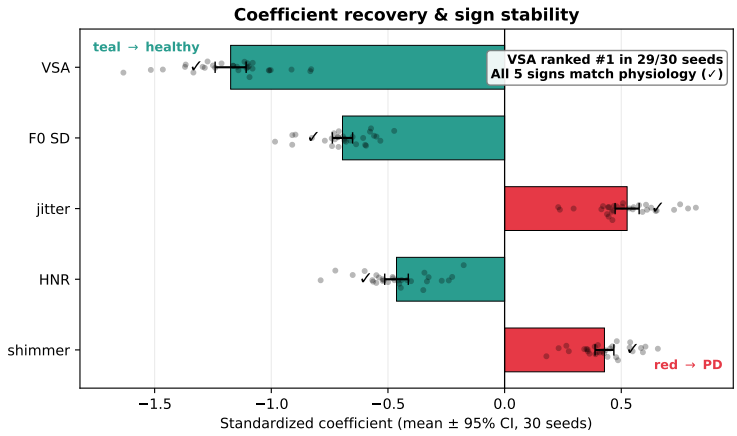

Fig. 4. Standardized coefficients of the federated model over 30 seeds (mean ± 95% CI; grey dots are per-seed values). Sign encodes direction (teal → healthy, red → PD); all five signs match physiology. VSA is the top-ranked feature in 29/30 seeds. The rank order among the phonatory descriptors is not stable and is not interpreted.

References

1. McMahan, B., Moore, E., Ramage, D., Hampson, S., Agüera y Arcas, B.: Communication-efficient learning of deep networks from decentralized data. In: Proc. 20th Int. Conf. on Artificial Intelligence and Statistics (AISTATS), PMLR, vol. 54, pp. 1273–1282 (2017)
2. Sheller, M.J., Edwards, B., Reina, G.A., et al.: Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci. Rep. **10**, 12598 (2020)
3. Pati, S., Baid, U., Edwards, B., et al.: Privacy preservation for federated learning in health care. Patterns **5**(7), 100974 (2024)
4. Tayebi Arasteh, S., Rios-Urrego, C.D., Nöth, E., Maier, A., et al.: Federated learning for secure development of AI models for Parkinson's disease detection using speech from different languages. In: Proc. Interspeech 2023, pp. 5003–5007 (2023)
5. Sarlas, A., Kalafatelis, A., Alexandridis, G., Kourtis, M.-A., Trakadas, P.: Exploring federated learning for speech-based Parkinson's disease detection. In: Proc. 18th Int. Conf. on Availability, Reliability and Security (ARES), pp. 1–8. ACM (2023)
6. Lee, H., Seo, D.: FedLC: Optimizing federated learning in non-IID data via label-wise clustering. IEEE Access **11**, 42082–42095 (2023)
7. Moro-Velázquez, L., Gómez-García, J.A., Arias-Londoño, J.D., Dehak, N., Godino-Llorente, J.I.: Advances in Parkinson's disease detection and assessment using voice and speech: a review of the articulatory and phonatory aspects. Biomed. Signal Process. Control **66**, 102418 (2021)
8. Skodda, S., Visser, W., Schlegel, U.: Vowel articulation in Parkinson's disease. J. Voice **25**(4), 467–472 (2011)

9. Rusz, J., Cmejla, R., Tykalová, T., et al.: Imprecise vowel articulation as a potential early marker of Parkinson's disease: effect of speaking task. J. Acoust. Soc. Am. **134**(3), 2171–2181 (2013)
10. Sapir, S., Ramig, L.O., Spielman, J.L., Fox, C.: Formant centralization ratio: a proposal for a new acoustic measure of dysarthric speech. J. Speech Lang. Hear. Res. **53**(1), 114–125 (2010)
11. Liu, Y., Penttilä, N., Ihalainen, T., et al.: Language-independent approach for automatic computation of vowel articulation features in dysarthric speech assessment. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 2228–2243 (2021)
12. Shen, M., Mortezaagha, P., Rahgozar, A.: Explainable artificial intelligence to diagnose early Parkinson's disease via voice analysis. Sci. Rep. **14**, 26398 (2024)
13. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 4765–4774 (2017)