When the Device Changes Mid-Trial: Speech-Based Digital Endpoint Robustness to On-Device Preprocessing Variability

Adrián Ortiz · Aggelos Ioannidis · Lisa Chiang · CEPHALGO SAS, Strasbourg, France · Correspondence: lisa.chiang@cephalgo.com

The Methodological Issue Being Addressed

Smartphone speech tasks are increasingly used as digital endpoints for remote longitudinal monitoring in neurodegenerative disease [1]. Studies follow participants over months to years, during which the recording device receives operating-system (OS) updates that can change on-device noise cancellation (NC), modifying the audio before it reaches the model.

Because the change is to the audio itself, it is upstream of every feature computed from it, handcrafted measures and learned embeddings alike. **The resulting shift is independent of clinical state, and it enters the data silently, becoming confounded with genuine longitudinal change.**

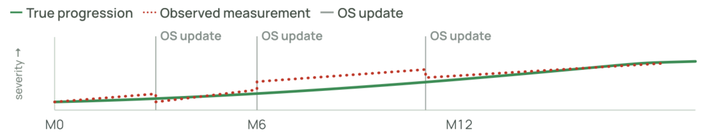

Figure 1. Confounding of OS updates with progression (schematic). A single feature stream cannot separate the two causes: true severity rises smoothly, while an OS update can add a step to the recorded measurement that mimics clinical change. The size and direction of the step vary, and not every update produces one.

Introduction (Aims)

We measure how much a simulated change in noise cancellation shifts a speech model's output, and whether a multi-stream representation is more stable than single-stream baselines.

Methods

Device manufacturers do not disclose their on-device noise cancellation, so we approximate an OS update with standard, widely used denoising methods applied to the same audio. Each recording receives one Gaussian-noise level (0.1 to 1.5% across the set), held fixed while denoised by each of three methods in turn, so only the denoising stage differs. Real deployed enhancement moves harmonics-to-noise ratio on dysphonic voice across six videoconferencing platforms [2]; the magnitude of a real OS update is left to future validation.

NC variants, successive OS updates over a trial timeline

v1 Spectral subtraction [3] **v2** Wiener filter [4] **v3** Wavelet denoising [5]
v1 → v2 1st update **v1 → v3** 2nd update **v1+v2 → v3** 2nd update, two configs seen

An independent external public corpus [6], processed by CEPHALGO: 28 PD, 22 age-matched controls; sustained vowels, syllable repetition and read passages; stratified 80/20 hold-out, no speaker overlap; 9 unseen test subjects (136 recordings), all metrics out-of-domain. Each model is trained on early denoiser variants (the simulated launch configuration), evaluated on a later, unseen variant.

Results

Multi-stream model

The model reads each recording through three parallel streams, chosen to differ in their sensitivity to the channel while each carries disease-related signal, combined into one prediction.

- **Linguistic:** Syntax, lexical diversity, pause and fluency.
- **Acoustic:** F0, jitter, shimmer, MFCCs, spectral shape, energy. Device sensitivity varies by measure [7, 8].
- **Self-supervised:** Pretrained embedding of the full audio, capturing signal beyond the named features.

Baselines pair a single representation, handcrafted acoustics or a self-supervised embedding (Wav2Vec2 [9], HuBERT [10], Whisper [11]), with an XGBoost classifier; the proposed model combines all streams. All models classify well in domain (raw→raw, AUC 0.96 to 1.00), so the comparison isolates robustness to the denoiser shift from a comparable in-domain starting point.

Reproducibility vs discrimination

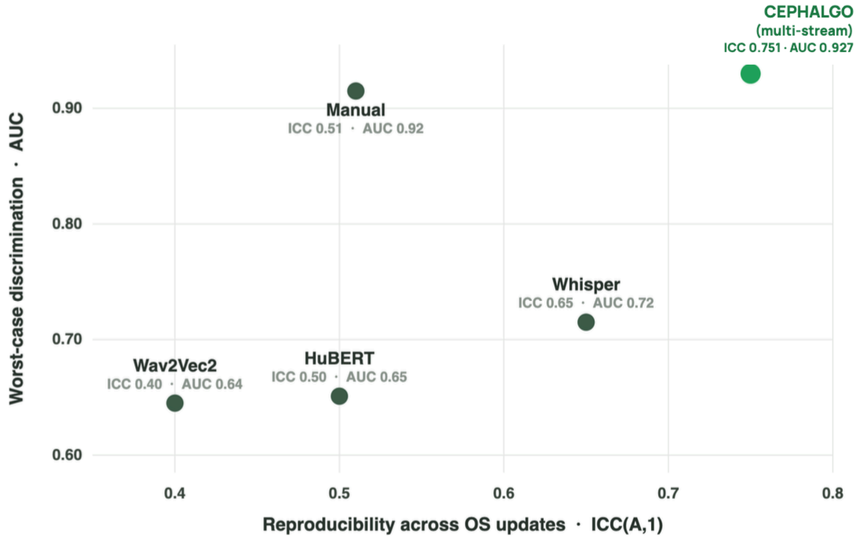

Figure 2. Only the multi-stream model occupies the top right. Reproducibility is highest for it (ICC 0.751, 2% NC share) and lowest for Wav2Vec2 and HuBERT (0.40, 0.50). The handcrafted stream reaches similar discrimination (AUC 0.92 vs the proposed model's 0.93) but is far less reproducible (ICC 0.51, 22% NC share).

Robustness distribution

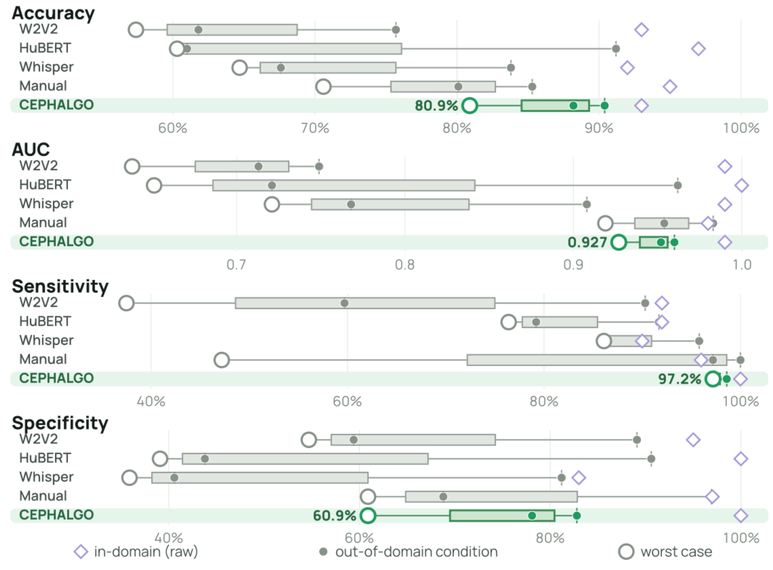

Figure 3. Green = proposed, grey = baselines. The multi-stream model holds worst-case sensitivity at 97.2% with no drop from in domain; its own worst-case specificity is 60.9%. The handcrafted stream's worst-case sensitivity falls from 100% to 47.2%, while its own worst-case specificity is also 60.9%, reached at a different condition from its worst sensitivity. Accuracy combines each condition's own sensitivity and specificity at the observed prevalence (72 PD recordings of 136).

Per-recording consistency

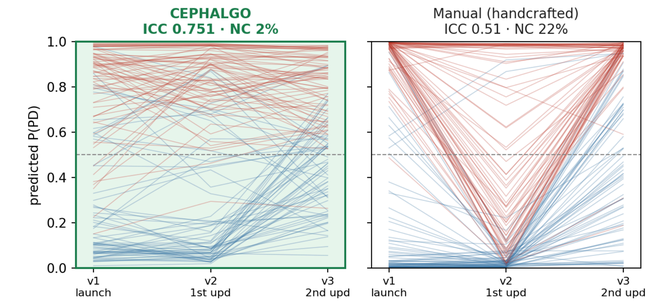

Figure 4. Each line is one recording, same audio through all three denoisers, model held fixed, so vertical movement comes from the denoiser alone. Proposed model (left): patient predictions hold position (median 0.89, 0.88, 0.84); controls rise at the second update (0.11 to 0.43), where worst-case specificity is lost. Handcrafted stream (right): patients collapse at the first update, crossing the controls, and recover at the second — its output follows the denoiser.

A reconstruction of a published feature panel shows the same effect

Panels of this kind vary by group; the measures used here are reported across [12–15]. One published combination of 18 moves under the same denoiser proxy.

LARGEST MOVERS, WORST-CASE CHANGE

Vowel space area	11.9%
F2 range upper	8.4%
F1 range upper	7.8%
Harmonicity range lower	7.8%
Second formant range	6.1%
Harmonicity range upper	6.0%
F1 range lower	5.6%
Articulation power	5.5%
Intensity variability, voiced	4.4%
Spectral shape modulation	4.3%
CPP	3.7%

Figure 5. Each of the 18 measures' worst-case change across the three post-update conditions, as a percentage of its launch value; largest movers shown.

Conclusions

1. A simulated change in on-device noise cancellation can move a speech model's output for a recording whose clinical state is unchanged, and the size of the effect depends on the model.
2. In this simulation the multi-stream representation combined low denoiser sensitivity, high reproducibility and the highest worst-case discrimination, while each single-representation baseline fell short on at least one, consistent with the combination of streams.
3. The same manipulation also moves a reconstruction of a published feature panel reported as a candidate endpoint. Because the re-derivation needs no OS metadata, this device-robust check can run on recordings already collected, alongside an existing single-stream endpoint.

Limits. The denoising methods approximate device NC; the magnitude of a real OS update is not disclosed and remains to be established. Where a panel measure's formulation is not disclosed, a standard acoustic definition was used. Worst-case specificity is the proposed model's weakest dimension; baselines vary in which metric degrades most. One corpus, one language, nine test speakers.

FUNDING

Co-funded by the European Union. European Innovation Council Accelerator, ADAPTE, Grant No. 190129251. CEPHALGO also participates in the Horizon Europe project OPADE under grant agreement No 101095436. Views expressed are those of the authors only and do not necessarily reflect those of the European Union or the granting authority.

ACKNOWLEDGEMENT

The authors thank the depositors of the public corpus. Open to retrospective and prospective collaboration.

DISCLOSURES

Adrián Ortiz, Aggelos Ioannidis and Lisa Chiang are employees of CEPHALGO SAS, which develops speech-derived digital measures. The work used a public third-party corpus and no commercial funding. One or more authors report potential conflicts which are described in the program.

Co-funded by the European Union

REFERENCES

[1] Lipsmeier, F. et al, Sci Rep, 12: 12081 (2022).
[2] Weerathunge, H.R. et al, J Speech Lang Hear Res, 64: 2586–2599 (2021). doi:10.1044/2021_JSLHR-20-00625.
[3] Boll, S., IEEE Trans ASSP, 27: 113–120 (1979).
[4] Lim, J.S. & Oppenheim, A.V., Proc IEEE, 67: 1586–1604 (1979).
[5] Donoho, D.L. & Johnstone, I.M., Biometrika, 81: 425–455 (1994).
[6] Dimauro, G. & Girardi, F., IEEE DataPort (2019). doi:10.21227/aw6b-tg17; cohort in Dimauro, G. et al, IEEE Access, 5: 22199–22208 (2017).
[7] Barsties v. Latoszek, B. et al, Am J Speech Lang Pathol (2025). doi:10.1044/2025_AJSLP-25-00140.
[8] Fahed, V.S. et al, J Voice, 39(2): 559.e1–559.e18 (2025).
[9] Baevski, A. et al, wav2vec 2.0, NeurIPS (2020).
[10] Hsu, W.-N. et al, IEEE/ACM TASLP, 29: 3451–3460 (2021).
[11] Radford, A. et al, arXiv:2212.04356 (2022).
[12] Kwasny, D. et al, speech measures in early Parkinson's disease, 18th AD/PD, Lisbon (2024).
[13] Kwasny, D. et al, vowel space metrics in early Parkinson's disease, AD/PD symposium, Vienna (2025).
[14] Rusz, J. et al, J Acoust Soc Am, 134(3): 2171–2181 (2013). doi:10.1121/1.4816541.
[15] Šimek, M. & Rusz, J., J Acoust Soc Am, 150(6): 4522 (2021). doi:10.1121/10.0009063.