Acoustic Representation Learning for Early Detection of Parkinson's Disease Through Vocal Expression and Speech Pattern Analysis

Cloe Luque
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
cloel@unh.edu

Te Koha Kenny
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
tekohakenny904@uab.edu

Millie Stratford
School of Computing, Clemson University, Clemson, SC, USA.
milliemail@clemson.edu

Abstract

Early detection of Parkinson's disease remains difficult because motor signs emerge after substantial dopaminergic loss and because speech changes are subtle, variable, and easily confounded by speaker identity, recording conditions, language, and task. Acoustic analysis has long suggested that hypophonia, monotonicity, imprecise articulation, and altered timing carry diagnostic information, yet handcrafted features often generalize poorly across cohorts and devices. This manuscript develops a methodological framework for acoustic representation learning that treats vocal expression and speech patterning as structured, multi-task signals rather than as isolated scalar biomarkers. The central contribution is a clinically constrained learning design that combines self-supervised spectrogram pretraining, supervised Parkinson's disease prediction, auxiliary speech tasks, domain-invariance penalties, calibrated uncertainty, and an abstention-aware decision rule. The framework specifies how waveform inputs are transformed into acoustic representations, how pathology-relevant variation can be separated from nuisance variation, and how representation quality should be evaluated with speaker-disjoint and site-held-out validation. Analytical attention is given to failure mechanisms, including shortcut learning from microphones or silence, label ambiguity in prodromal disease, medication-state effects, and the risk that excessive domain invariance removes clinically meaningful signal. The paper also defines an evaluation logic based on external validation, calibration, decision-curve analysis, and transparent reporting. The argument is that early vocal screening requires representations that are sensitive to early dysarthric change, stable across acquisition conditions, and linked to interpretable acoustic evidence. Such representations can support clinician-in-the-loop risk stratification while avoiding claims of autonomous diagnosis.

Keywords

Parkinson's disease, acoustic representation learning, speech biomarkers, self-supervised learning, domain generalization, clinical decision support

1. Introduction

Parkinson's disease is a progressive neurodegenerative disorder whose clinical diagnosis depends on the presence of bradykinesia, rest tremor, rigidity, and supportive features, with a definite diagnosis historically requiring postmortem confirmation [1]. Clinical criteria have improved diagnostic accuracy, and the Movement Disorder Society criteria formalize supportive and red-flag features, but the early phase remains ambiguous because several conditions mimic Parkinsonism and because symptom onset is insidious [2]. The practical consequence is that many patients are identified only after motor manifestations are well established. A screening tool that could flag subtle vocal and speech changes during routine interaction would therefore be valuable, not as a replacement for neurological assessment, but as a triage and monitoring aid that encourages timely specialist evaluation.

Speech is an attractive signal for this purpose. Phonation, articulation, resonance, respiration, and prosody depend on integrated basal ganglia, cerebellar, cortical, and brainstem control, and Parkinsonian dysarthria commonly involves reduced vocal intensity, restricted pitch range, imprecise consonant production, vowel centralization, and altered pause and timing patterns [3], [4]. These changes can be measured noninvasively from sustained vowels, diadochokinetic tasks, reading passages, and spontaneous speech. Early work using sustained phonation showed that dysphonia measures such as jitter, shimmer, harmonic-to-noise ratio, and recurrence-based descriptors can distinguish Parkinsonian voices from control voices and can support telemonitoring [5]. Subsequent work extended acoustic analysis to disease progression and to cross-linguistic running speech [6], [7]. The existence of these associations does not, however, imply that a single acoustic feature is a reliable diagnostic marker. Speech is a high-dimensional expressive behavior, and Parkinsonian changes overlap with aging, depression, respiratory disease, medication effects, and normal between-speaker variation.

Most existing computational studies follow a feature-engineering paradigm: acoustic descriptors are extracted from a small set of speech tasks, feature selection is performed, and a classifier is trained on a modest cohort. This approach has produced useful evidence, but it has three limitations. First, handcrafted features encode assumptions about which acoustic dimensions matter, and those assumptions may not transport across languages, recording devices, and speaking styles. Second, classification accuracy on a single corpus can reflect corpus-specific shortcuts, such as microphone characteristics, background noise, or speaker identity, rather than disease-related dysarthria. Third, binary classification alone does not specify how a model should behave when evidence is weak, when a recording is corrupted, or when a patient lies near a diagnostic boundary. These limitations motivate representation learning: instead of choosing all acoustic descriptors in advance, a model learns a mapping from raw or spectral speech to a latent space in which clinically relevant variation is preserved and nuisance variation is suppressed.

The central contribution of this manuscript is a clinically constrained acoustic representation learning framework for early Parkinson's disease detection from vocal expression and speech patterns. The framework combines self-supervised masked spectrogram modeling, supervised disease prediction, auxiliary speech tasks, domain-invariance regularization, uncertainty estimation, and an abstention-aware decision rule. It is not presented as a completed clinical study; it is a methodological and analytical treatment of how such a system should be formulated, where its assumptions are vulnerable, and how its claims should be evaluated. The argument is that representation learning can improve early

vocal screening only if it is designed around clinical endpoints, explicit confounding control, and external validation. A model that learns a compact representation without addressing acquisition domain, task effects, and label uncertainty will likely reproduce the weaknesses of earlier feature-based classifiers rather than overcome them.

The paper is organized as follows. Section 2 reviews the clinical and acoustic basis of Parkinsonian speech change and identifies sources of variability. Section 3 formulates the learning problem, including inputs, targets, and representation objectives. Section 4 develops the architecture and training strategy, with attention to self-supervision, auxiliary tasks, and domain invariance. Section 5 presents an evaluation and validation protocol that distinguishes methodological proposal from observed evidence. Section 6 discusses technical failure modes and boundary conditions. Section 7 considers clinical translation and data stewardship, and Section 8 concludes.

2. Clinical and Acoustic Background

2.1 Vocal manifestations and early disease

Parkinsonian dysarthria is traditionally described as hypokinetic. Reduced amplitude of movement affects respiratory support, laryngeal tension, and articulatory excursions, producing lowered vocal intensity, reduced fundamental-frequency variability, and blurred consonant contrasts. Perceptual studies report monotone pitch, short rushes of speech, imprecise consonants, and inappropriate silences, although individual profiles vary considerably [4]. Acoustic studies have quantified several of these dimensions: fundamental frequency and its perturbation, amplitude perturbation, harmonic-to-noise ratio, vowel formant centralization, voice onset time, consonant spectral moments, articulation rate, and pause distribution [3], [7]. The appeal for early detection is that speech production is sensitive to neuromotor control and can be sampled repeatedly with low burden.

The clinical interpretation of these measures requires caution. Acoustic changes are not pathognomonic. Reduced pitch range can occur in depression, fatigue, and laryngeal pathology; increased jitter can follow vocal fold lesions or poor recording quality; consonant imprecision can reflect dialect, dentition, or cognitive-linguistic factors. Even within Parkinson's disease, speech signs vary by phenotype, medication state, disease duration, and task. Sustained vowel phonation isolates laryngeal stability and is easy to standardize, but it omits the articulatory and prosodic coordination that running speech engages. Reading tasks improve comparability but introduce literacy and language effects. Spontaneous speech is ecologically valid but difficult to segment and control. A representation learning approach should therefore treat task identity as an explicit variable, not as an interchangeable source of data.

The mechanistic rationale for early vocal analysis is plausible but indirect. Basal ganglia dysfunction disrupts the scaling and timing of movement, and speech is a timed sensorimotor behavior with rapid sequential transitions. However, the presence of an acoustic difference in a clinical cohort does not establish that the difference is an early disease marker, a compensatory response, or a medication effect. The framework developed here treats acoustic representations as statistical correlates that must be validated against clinical adjudication and longitudinal outcome, not as direct readouts of dopaminergic loss.

2.2 Sources of variability and confounding

Speech signals carry at least four overlapping sources of variation: linguistic content, speaker identity and anatomy, expressive style and task, and recording channel. Parkinsonian dysarthria is a fifth source that may be entangled with all of them. Speaker identity contributes vocal tract length, habitual pitch, dialect, and idiosyncratic pronunciation. Task contributes speaking rate, formality, and phonetic inventory. Channel contributes microphone frequency response, compression, noise reduction, and room acoustics. If a classifier learns that one clinic uses a particular device or that Parkinsonian participants were recorded in a quieter room, it may exploit acquisition artifacts rather than disease biology.

Cross-linguistic work has shown that running speech can yield disease-related information across languages, but it also demonstrates that models trained in one language or recording condition do not automatically transfer to another [7]. The problem is not merely technical nuisance. Early disease itself is heterogeneous: tremor-dominant and akinetic-rigid presentations may differ in speech involvement, and cognitive or affective comorbidities can alter expressive prosody. A representation that is perfectly invariant to all speaker and task variation would be undesirable because some of that variation may carry clinical signal. The design goal is selective invariance: suppress variation that is irrelevant or misleading for the clinical question while preserving variation that plausibly reflects neuromotor dysfunction.

Work on style transfer for out-of-domain singing voice synthesis is relevant by analogy rather than by direct evidence [8]. That line of research addresses how timbre and expressive style can be manipulated while preserving linguistic content in synthetic voices. For clinical acoustics, the analogy is cautionary: vocal style and speaker timbre are powerful, separable factors in generated speech, and a model trained to predict disease can mistake stylistic or speaker-specific factors for pathology if the cohort is imbalanced. The task is therefore not to remove all expressive variation, but to identify which dimensions should be treated as nuisance for a given clinical endpoint and which should remain available for interpretation.

3. Problem Formulation and Learning Framework

3.1 Inputs, clinical targets, and task structure

Consider a dataset of participants indexed by i. Each participant contributes one or more recording sessions s, each session contains speech tasks k, and each task yields utterances m. The waveform for an utterance is denoted $a_{i,s,k,m}(t)$. The first preprocessing step converts the waveform into a time-frequency representation suitable for a neural encoder. A conventional choice is a short-time Fourier transform followed by log compression:

$$X = log(1 + |STFT(a)|^2)$$

Here a is the waveform, STFT denotes the short-time Fourier transform, and X is a nonnegative log-power spectrogram whose rows index time frames and columns index frequency bins. Log compression reduces the dynamic range of speech and makes the representation less sensitive to absolute gain, although it does not remove channel effects. Alternative front ends, such as log-mel filterbanks or learned convolutional front ends, can be substituted without changing the logic of the framework; the choice should be justified by validation rather than assumed.

The clinical target must be defined before model development. A binary label y_i can indicate whether participant i meets Movement Disorder Society criteria for Parkinson's disease at a prespecified assessment and has a disease duration below a prespecified early-disease threshold [2]. A more informative target is time to specialist-confirmed diagnosis,

which can be modeled as a survival outcome with censoring for participants not diagnosed during follow-up. The binary target is easier to evaluate but discards temporal information and can blur prevalent and incident disease. The survival target is closer to the early-detection question but requires larger cohorts and careful handling of censoring. A defensible framework can include both: a primary binary endpoint for screening performance and a secondary time-to-event endpoint for diagnostic delay.

The input representation should be pooled across utterances and tasks without erasing task identity. A simple formulation learns an utterance-level representation $z_{i,s,k,m}$ and aggregates it:

$$z = Pool(h\theta(X))$$

Here hθ is an encoder parameterized by θ, X is the spectrogram, and Pool is an aggregation operator over time frames and, optionally, over utterances and tasks. The pooled vector z is the representation used for clinical prediction. The pooling operation can be attention-based, mean pooling, or a learned query-based aggregation. Mean pooling is stable but may dilute short dysarthric events; attention pooling can highlight informative segments but requires regularization to avoid overfitting to sparse artifacts. Because different speech tasks probe different mechanisms, pooling should either retain a task embedding or use a multi-task architecture in which task identity is an explicit input.

3.2 Self-supervised representation learning

Supervised labels for early Parkinson's disease are expensive and noisy. Unlabeled speech, including speech from participants who never develop Parkinson's disease, is abundant. Self-supervised learning can exploit this asymmetry by training the encoder to predict masked portions of the spectrogram or to distinguish temporally related representations. Masked spectrogram modeling is attractive because it encourages the encoder to capture formant transitions, voicing continuity, and prosodic structure without requiring phoneme labels. The masked prediction loss can be written as:

$$Lssl = E_M \sum_\{t \in M\} ||x_t - g\varphi(h\theta(\tilde{X}))_t||_2^2$$

Here M is a random set of masked time-frequency positions, $\tilde{X}$ is the spectrogram with those positions masked, x_t is the target frame or patch at position t, gφ is a decoder parameterized by φ, and hθ is the encoder. The expectation is over masking masks and training examples. The loss is not a clinical objective; it is a representation-learning objective that shapes the encoder before or alongside supervised training. Its value depends on whether the learned features align with speech properties that are relevant to dysarthria. A model can reconstruct spectrograms accurately while ignoring subtle timing cues, so masked prediction should be complemented by tasks that emphasize articulation, prosody, and voice quality.

3.3 Supervised and auxiliary objectives

The clinical objective should be a proper scoring rule, such as binary cross-entropy for y_i, with class weighting or stratification only when justified by the sampling design. The total training objective can combine clinical prediction, self-supervision, auxiliary speech tasks, and domain regularization:

$$L = LPD + \lambda_1 Lssl + \lambda_2 Laux + \lambda_3 Ldom$$

Here LPD is the clinical prediction loss, Lssl is the self-supervised loss in Equation (3), Laux is an auxiliary speech-task loss, and Ldom is a domain-invariance penalty. The

coefficients λ_1, λ_2, and λ_3 are nonnegative weights selected by nested validation, not fixed constants. The auxiliary task could be phoneme classification from forced alignment, articulatory feature prediction, speech activity detection, or prosodic boundary prediction. Its role is to bias the representation toward speech structure that is likely to be stable across speakers. The auxiliary task must not leak the clinical label. If phoneme labels are generated by an automatic aligner that performs differently on dysarthric speech, the auxiliary loss may introduce bias rather than reduce it.

The domain penalty requires a definition of domain. Domains may be recording sites, microphone types, languages, sexes, age bands, or speaking tasks. A distributional penalty can align representations across domains:

Here A and B are two domains, nA and nB are their sample sizes, z_i is the representation for participant i, φ is a feature map into a reproducing kernel Hilbert space H, and the squared norm is the squared maximum mean discrepancy. The penalty encourages similar representation distributions across domains. It is a proposed regularizer, not a guarantee of causal invariance. If disease prevalence differs across domains, forcing distributional alignment can remove disease-related signal. For that reason, domain penalties should be applied within strata or with stratified losses, and their effect should be tested by ablation.

4. Acoustic Representation Learning Architecture

4.1 Encoder design for vocal expression and speech patterns

The encoder should operate on time-frequency structure and preserve temporal resolution. Convolutional networks are effective at capturing local spectro-temporal patterns, such as formant transitions and voice-onset bursts, while transformer layers can model longer-range prosodic dependencies. A hybrid design with convolutional stem, residual blocks, and self-attention is plausible. The convolutional stem reduces frequency variation and handles phase-insensitive spectral patterns; the attention layers model pauses, rhythm, and intonation contours over longer windows. The model should receive task identity as an embedding so that sustained phonation and running speech are not pooled into a single undifferentiated distribution. It should also receive recording metadata when available, not because metadata should be used for prediction, but because it can be used to define domains and to diagnose shortcut learning.

The representation z should be evaluated at multiple levels. A participant-level vector supports clinical prediction. Frame-level or segment-level representations support interpretability and can be probed for acoustic attributes such as pitch range, vowel space, and pause duration. The framework should not rely on a single opaque vector if the clinical claim is that the model detects dysarthria. If the model's decision cannot be related to known speech dimensions, the evidence for clinical validity is weaker. Probing does not prove causality, but it helps distinguish a model that uses pathological speech variation from one that uses microphone artifacts or silence.

Self-supervised pretraining can use speech from the target clinical population and from external corpora. Pretraining on large general speech corpora may learn generic phonetic structure, but it may also learn recording conditions that differ from the clinical setting. Pretraining on unlabeled clinical speech may be more relevant, but privacy constraints and cohort size limit availability. A practical strategy is staged pretraining: first learn general speech structure from broad corpora, then adapt to the clinical domain with masked prediction on unlabeled recordings from the target sites. The adaptation stage should be evaluated for

catastrophic forgetting and for domain-specific overfitting. Models such as wav2vec 2.0 and HuBERT demonstrate that self-supervised speech representations can capture phonetic and acoustic structure without labels [9], [10], and contrastive predictive coding provides an earlier framework for learning useful temporal representations [11]. These methods are not clinically validated for Parkinson's disease; they are architectural and methodological precedents.

4.2 Multi-task learning and selective invariance

The multi-task objective in Equation (4) creates a tension between clinical sensitivity and domain invariance. Early dysarthria may appear as increased variability in timing or pitch, but domain-invariance penalties can suppress variability if it is correlated with speaker or recording condition. The solution is not to remove all variability but to define which variation should be preserved. One approach is to use a domain classifier with a gradient reversal layer, so that the encoder is penalized for making domain identity predictable [12]. Another approach is invariant risk minimization, which seeks predictors whose risk is stable across environments [13]. Both approaches assume that domains are known and that the training environments are representative. In clinical speech, domains are often partially observed and confounded with disease prevalence. A model trained to be invariant across sites may become invariant to disease severity if one site recruits more advanced patients.

A more defensible design is to combine domain-invariance penalties with explicit nuisance augmentation. Channel augmentation can simulate microphone frequency response, additive noise, and compression. Speaker augmentation can alter pitch and speaking rate within physiologically plausible ranges. Task augmentation can mix or mask segments. These augmentations should be clinically plausible; extreme pitch shifting may create artifacts that the encoder learns to ignore rather than learning true speaker invariance. Augmentation is not a substitute for external validation, but it can reduce the model's reliance on superficial cues.

Auxiliary tasks should be chosen to encourage speech-relevant features. Phoneme prediction encourages articulatory contrast; prosodic boundary prediction encourages timing and intonation; voice-quality prediction encourages laryngeal stability. However, if auxiliary labels are generated from the same acoustic features that the clinical model uses, the auxiliary task may reinforce existing biases. A useful auxiliary task should require the encoder to represent a different aspect of speech than the clinical label. For example, predicting phoneme identity from a masked spectrogram encourages the encoder to represent formant transitions, while predicting pause boundaries encourages temporal structure. The auxiliary loss weight should be tuned so that clinical performance is not sacrificed for auxiliary accuracy.

4.3 Uncertainty and abstention

A clinical screening model should not output a probability without an uncertainty estimate. Two sources of uncertainty matter. Aleatoric uncertainty arises from ambiguous recordings, overlapping speech, or label noise. Epistemic uncertainty arises from model uncertainty about participants or recording conditions not represented in training. For a binary prediction, a simple calibrated probability is:

$$p_i = \sigma\left(w^T z_i + b\right)$$

Here p_i is the predicted probability that participant i meets the early Parkinson's disease endpoint, z_i is the pooled representation, w is a weight vector, b is a bias term, and σ is the logistic function. Calibration should be assessed on held-out data because modern neural

networks can be overconfident even when accurate [14]. Calibration methods such as temperature scaling or isotonic regression can be applied, but they must be fit on data separate from the test set.

A predictive entropy measure can flag ambiguous cases:

$$u_i = -[p_i \log p_i + (1 - p_i)\log(1 - p_i)]$$

Here u_i is the binary entropy of the predicted probability. High entropy indicates that the model is close to a decision boundary; it does not by itself indicate that the recording is poor. Epistemic uncertainty can be estimated with deep ensembles or Monte Carlo dropout, although these methods add computational cost and require validation [15]. The decision rule can incorporate both probability and uncertainty:

$$d_i = 1 \; if \; p_i \geq \tau \; and \; u_i \leq u_{max};\; 0 \; if \; p_i < \tau \; and \; u_i \leq u_{max};\; abstain \; otherwise$$

Here d_i is the screening decision, τ is the probability threshold, and u_{max} is the maximum admissible entropy. The abstention zone is not a failure of the model; it is a safety mechanism that routes uncertain cases to specialist review or repeat recording. The thresholds τ and u_{max} should be selected using clinical utility analysis rather than accuracy alone. Decision curve analysis can quantify net benefit across a range of threshold probabilities and make the trade-off between false positives and false negatives explicit [16].

5. Evaluation and Validation Logic

5.1 Study design and endpoint adjudication

The evaluation of an acoustic representation learning model for early Parkinson's disease must be designed around the clinical question. A cross-sectional comparison of diagnosed patients and healthy controls answers whether speech differs between groups, but it does not answer whether the model detects disease early. A stronger design enrolls participants at risk, such as those with REM sleep behavior disorder, hyposmia, subtle motor signs, or family history, and follows them longitudinally to specialist-confirmed diagnosis. The primary endpoint should be the first diagnosis by a movement disorder specialist using established criteria [2]. Adjudicators should be blinded to model outputs and to acoustic features. A secondary endpoint can be time from first recording to diagnosis, analyzed with survival methods that account for censoring.

The study should include participants with conditions that mimic Parkinson's disease, because a screening model that distinguishes only healthy controls from advanced Parkinson's disease will overstate clinical value. Essential tremor, drug-induced parkinsonism, vascular parkinsonism, and atypical Parkinsonian syndromes are relevant comparators. The model's error profile on these conditions is part of its clinical validity, not an afterthought. Clinical severity should be measured with the Movement Disorder Society Unified Parkinson's Disease Rating Scale, and analyses should account for medication state and motor subtype [19]. Speech recordings should be collected in a standardized protocol that includes sustained phonation, diadochokinetic tasks, reading, and spontaneous speech, with metadata on device, environment, and task order.

5.2 Speaker-disjoint and site-held-out validation

Random train-test splits at the recording level are inadequate because recordings from the same participant can appear in both training and test sets. Speaker-disjoint splits ensure that the model is evaluated on unseen voices. Site-held-out validation ensures that the model is

evaluated on unseen recording environments. Both are necessary. If the dataset contains multiple languages, language-held-out validation tests cross-linguistic transport. If the dataset contains longitudinal recordings, time-held-out validation tests whether the model predicts future diagnosis rather than memorizing prevalent disease. Nested cross-validation should be used for hyperparameter selection, with all preprocessing, feature selection, and calibration fit inside the training folds. The test set should be touched only once for the final evaluation.

External validation is the strongest test. A model developed on one cohort should be evaluated on an independent cohort with a different recruitment strategy, device, and language when possible. External validation should report discrimination, calibration, and clinical utility. Discrimination can be summarized by area under the receiver operating characteristic curve and area under the precision-recall curve, but these metrics do not capture calibration or decision consequences. Calibration should be reported with calibration plots or calibration slope and intercept. Clinical utility should be reported with decision curve analysis or net benefit [16]. Reporting should follow the TRIPOD statement for prediction models, including participant flow, predictor definitions, missing data, and model specification [17].

5.3 Baseline comparisons and ablation

The representation learning framework should be compared with credible baselines. Handcrafted acoustic features such as jitter, shimmer, harmonic-to-noise ratio, formant centralization, pause duration, and speech rate remain important benchmarks because they encode decades of acoustic research and are interpretable [3], [5], [6]. A linear classifier on these features is a strong test of whether representation learning adds value. A supervised neural network trained from scratch on the same spectrograms tests whether self-supervised pretraining helps. A self-supervised model without clinical fine-tuning tests whether the representation alone captures disease-related variation. A model with domain-invariance penalties should be compared with a model without them to determine whether invariance improves external validation or removes signal.

Ablation should also examine task composition. Sustained vowels and running speech carry different information, and a model that performs well only when both tasks are present may be unsuitable for settings where only one task is feasible. Auxiliary task gains should be evaluated by removing each auxiliary loss in turn. Uncertainty should be evaluated by measuring the fraction of abstentions and the error rate among non-abstained cases. A model that improves accuracy by abstaining on a large fraction of cases may have limited utility; a model that abstains only on corrupted recordings may be more useful. These analyses require prespecified metrics and transparent reporting.

5.4 Interpretability and acoustic probing

Interpretability should be treated as part of validation, not as a post hoc visualization. Probe classifiers can test whether the learned representation encodes known acoustic dimensions such as fundamental frequency range, vowel dispersion, consonant spectral moments, and pause distribution. If the model's representation does not encode these dimensions, its clinical claim is weaker, because the known speech phenotype is not recoverable. If the representation encodes them but the clinical classifier ignores them, the model may be using shortcut features. Attention maps and saliency methods can suggest which time-frequency regions drive prediction, but they are not causal explanations. They should be compared with acoustic measurements and with clinician judgments.

A useful interpretability test is counterfactual augmentation: if the model's prediction is driven by pitch instability, then augmenting pitch stability should change the prediction in a predictable direction. If the model's prediction is driven by microphone noise, then channel augmentation should reduce the prediction's dependence on noise. These tests do not prove clinical validity, but they can falsify specific shortcut hypotheses. The representation should also be tested for demographic bias: performance should be stratified by sex, age, language, and recording device. A model that performs well overall but poorly in older women or in a minority language is not ready for clinical use.

6. Technical Discussion and Boundary Conditions

The proposed framework has several boundary conditions. First, the clinical endpoint is not the same as the acoustic endpoint. A model can learn to predict Parkinson's disease from speech because speech changes with age, medication, or general health, or because recruitment sites differ. Association between representation and disease is necessary but not sufficient for early detection. Longitudinal validation and mimic controls are required to distinguish a disease marker from a correlated feature. Second, self-supervised learning does not guarantee clinically relevant features. Masked spectrogram modeling can learn to reconstruct background noise, room reverberation, or codec artifacts. If these factors are correlated with disease in the training data, the model will use them. Domain-invariance penalties and augmentation can reduce the risk, but they cannot eliminate confounding that is unmeasured.

Third, domain invariance can be harmful if it is applied too aggressively. Parkinsonian speech varies by phenotype and severity. If the model is trained to be invariant across sites that differ in disease prevalence or severity, it may learn to ignore the very features that distinguish early disease. Selective invariance requires a clear definition of nuisance and a validation strategy that tests whether invariance improves transport without reducing sensitivity. Fourth, uncertainty estimates are only useful if they are calibrated and if the abstention thresholds are set by clinical stakeholders. An entropy threshold that is statistically convenient may be clinically unacceptable if it withholds too many high-risk patients or if it sends too many low-risk patients to specialists.

Fifth, speech tasks are not interchangeable. Sustained phonation provides a stable measure of laryngeal function but misses articulation and prosody. Reading passages provide controlled phonetic content but introduce literacy effects. Spontaneous speech is ecologically valid but difficult to standardize. A multi-task model can exploit complementary information, but it can also learn task-specific shortcuts. The model should report performance by task and should be evaluated in settings where the full task battery is unavailable. A screening tool that requires a 20-minute recording may be less useful than a shorter tool with slightly lower discrimination, depending on the deployment context.

Sixth, label noise in early Parkinson's disease is substantial. Clinical diagnosis can change over time, especially in the first years after symptom onset. A participant labeled as a control at baseline may later be diagnosed with Parkinson's disease. If the model is trained on baseline labels without follow-up, it may be penalized for correctly detecting prodromal disease. Longitudinal adjudication and survival analysis can address this problem, but they require long follow-up. A practical alternative is to treat uncertain diagnoses as censored or as a separate category, and to evaluate the model under different label assumptions. This is not a technical detail; it affects the interpretation of every performance metric.

Seventh, the acoustic representation must be linked to vocal expression, not only to speech content. Prosody, voice quality, and timing are expressive dimensions that may be altered in early Parkinson's disease. A model that learns only phoneme identity may miss these dimensions. Auxiliary tasks should therefore include prosodic and voice-quality targets when labels are available. If labels are unavailable, self-supervised objectives can be designed to predict prosodic boundaries, pitch contours, or energy envelopes. These objectives are not disease-specific, but they bias the representation toward expressive speech properties that are clinically plausible.

Seventh, the model's decision rule should be evaluated in the context of the clinical pathway. A screening aid that flags high-risk participants for specialist review has different costs and benefits from a monitoring aid that tracks disease progression. The threshold τ and abstention bound u_{max} should be chosen with clinicians, patients, and health system constraints. Decision curve analysis can make these trade-offs explicit, but it cannot resolve them. The framework therefore should not be presented as a diagnostic device. Its proper claim is that a well-validated representation can improve risk stratification and reduce diagnostic delay when embedded in a clinical pathway.

7. Clinical Translation and Data Stewardship

Clinical translation of acoustic screening requires attention to privacy and voice biometrics. Voice recordings can identify speakers, reveal emotional state, and contain background conversations. Data governance should include informed consent for secondary use, de-identification that preserves acoustic properties needed for analysis, and secure storage. Models trained on clinical speech should be tested for membership inference and speaker identification risks. If a model is deployed as a smartphone application, the on-device processing and data transmission policy should be transparent. These concerns are not unique to Parkinson's disease, but they are especially important because voice is a persistent biometric [20].

Bias and equity require explicit evaluation. Parkinson's disease prevalence and speech characteristics differ by age, sex, language, and comorbidity. Recording devices and digital literacy affect data quality. A model developed on high-quality laboratory recordings may fail in primary care or home settings. External validation should include diverse populations and low-resource languages. If performance is lower in a subgroup, the model should not be deployed until the cause is understood. The framework's domain-invariance components may help, but they are not a substitute for inclusive data collection.

Clinical implementation should be clinician-in-the-loop. A screening model can prioritize patients for neurological assessment, but it should not diagnose Parkinson's disease or initiate treatment. The output should include a probability, an uncertainty indicator, and a concise acoustic summary that a clinician can interpret. The summary might report pitch range, vowel space, pause distribution, and voice stability relative to normative data, but these values must be validated and presented without implying diagnostic certainty. The model should be monitored after deployment for drift in recording devices, population characteristics, and disease prevalence.

8. Conclusion

Early detection of Parkinson's disease through vocal expression and speech pattern analysis is plausible because speech is sensitive to neuromotor control and because acoustic recording is noninvasive, repeatable, and scalable. The field does not lack acoustic correlates;

it lacks representations that are sensitive to early dysarthric change, stable across acquisition conditions, and evaluated with clinical rigor. This manuscript has developed a methodological framework for acoustic representation learning that combines self-supervised spectrogram pretraining, supervised disease prediction, auxiliary speech tasks, domain-invariance regularization, calibrated uncertainty, and abstention-aware decision making. The framework defines inputs and endpoints, specifies a multi-task objective, and explains how representation learning can be evaluated with speaker-disjoint and site-held-out validation.

The central analytical claim is that selective invariance is necessary. A model that is invariant to all speaker, task, and channel variation will discard clinically meaningful expression; a model that is invariant to nothing will learn shortcut features. The design challenge is to identify the nuisance dimensions that should be suppressed and the expressive dimensions that should be preserved. This requires clinical input, acoustic probing, ablation, external validation, and longitudinal outcome data. The framework also requires calibrated uncertainty and a decision rule that allows abstention when evidence is weak. These features are not add-ons; they are part of the clinical validity of a screening model.

The next step for this line of research is not to build a larger classifier but to build a better evidentiary chain. That chain begins with a prespecified clinical endpoint, continues through task-standardized recording, speaker-disjoint and site-held-out validation, comparison with handcrafted acoustic baselines, calibration and decision-curve analysis, and ends with prospective evaluation in a clinical pathway. Representation learning can contribute to that chain by learning acoustic features that are difficult to specify by hand, but it cannot replace clinical adjudication or external validation. If developed with these constraints, acoustic representation learning can support earlier specialist referral and more equitable monitoring while remaining honest about the difference between a statistical correlate and a clinical diagnosis.

References

1. Jankovic, J. (2008). Parkinson's disease: Clinical features and diagnosis. Journal of Neurology, Neurosurgery & Psychiatry, 79(4), 368–376. https://doi.org/10.1136/jnnp.2007.131045

2. Postuma, R. B., Berg, D., Stern, M., Poewe, W., Olanow, C. W., Oertel, W., Obeso, J., Marek, K., Litvan, I., Lang, A. E., Halliday, G., Goetz, C. G., Gasser, T., Dubois, B., Chan, P., Bloem, B. R., Adler, C. H., & Deuschl, G. (2015). MDS clinical diagnostic criteria for Parkinson's disease. Movement Disorders, 30(12), 1591–1601. https://doi.org/10.1002/mds.26424

3. Rusz, J., Cmejla, R., Ruzickova, H., & Ruzicka, E. (2011). Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. The Journal of the Acoustical Society of America, 129(1), 350–367. https://doi.org/10.1121/1.3514381

4. Sapir, S., Ramig, L. O., & Fox, C. M. (2008). Speech and swallowing disorders in Parkinson disease. Current Opinion in Otolaryngology & Head and Neck Surgery, 16(3), 205–210. https://doi.org/10.1097/MOO.0b013e3282febd3a

5. Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE

Transactions on Biomedical Engineering, 56(4), 1015–1022. https://doi.org/10.1109/TBME.2008.2005954

6. Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2010). Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. IEEE Transactions on Biomedical Engineering, 57(4), 884–893. https://doi.org/10.1109/TBME.2009.2036000

7. Orozco-Arroyave, J. R., Hönig, F., Arias-Londoño, J. D., Vargas-Bonilla, J. F., Daqrouq, K., Skodda, S., Rusz, J., & Nöth, E. (2016). Automatic detection of Parkinson's disease in running speech spoken in three different languages. The Journal of the Acoustical Society of America, 139(1), 481–500. https://doi.org/10.1121/1.4939739

8. Zhang, Y., Huang, R., Li, R., He, J., Xia, Y., Chen, F., ... & Zhao, Z. (2024, March). Stylesinger: Style transfer for out-of-domain singing voice synthesis. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 17, pp. 19597-19605).

9. Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Advances in Neural Information Processing Systems (Vol. 33, pp. 12449–12460).

10. Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291

11. van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

12. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(59), 1–35.

13. Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. arXiv preprint arXiv:1907.02893.

14. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 1321–1330).

15. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems (Vol. 30, pp. 5574–5584).

16. Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. Medical Decision Making, 26(6), 565–574. https://doi.org/10.1177/0272989X06295361

17. Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. Annals of Internal Medicine, 162(1), 55–63. https://doi.org/10.7326/M14-0697

18. Rusz, J., Cmejla, R., Tykalova, T., Ruzickova, H., Klempir, J., Majerova, V., & Ruzicka, E. (2013). Imprecise vowel articulation as a potential early marker of Parkinson's disease:

Effect of speaking task. The Journal of the Acoustical Society of America, 134(3), 2171–2181. https://doi.org/10.1121/1.4816541

19. Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M. B., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A. E., Lees, A., Lehericy, S., Lewitt, P. A., Nyenhuis, D., ... & LaPelle, N. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. Movement Disorders, 23(15), 2129–2170. https://doi.org/10.1002/mds.22340

20. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7