Acoustic Biomarker Extraction for Early Detection of Neurological Disorders Using Generative Speech Representations

Rune Mohammed
Department of Computer Science, George Mason University, Fairfax, VA, USA.
mohammed1975@gmu.edu

Ada Heinz
School of Computing, Clemson University, Clemson, SC, USA.
heinz2002@clemson.edu

Jinthe Lagerweij
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
contactjinthe@colostate.edu

Eliza Schenkel
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
eliza.work@binghamton.edu

Abstract

Early neurological disorders can alter speech before routine clinical diagnosis, yet acoustic biomarkers remain difficult to transfer across speakers, languages, recording conditions, and disease stages. Conventional features such as jitter, shimmer, formant trajectories, pause duration, and articulation rate are interpretable but brittle, while self-supervised speech representations are information-rich but not directly equivalent to measurable biomarkers. This manuscript develops a methodological framework for acoustic biomarker extraction that couples self-supervised speech encoders with a generative latent model, counterfactual resynthesis, calibrated risk estimation, and longitudinal scoring. The central contribution is a representation-level construction in which speech is decomposed into content, speaker style, and a latent neuromotor state; biomarkers are defined as segment-specific deviations from a conditional normative manifold rather than as raw classifier scores. Counterfactual normalization preserves linguistic content and speaker identity while replacing the estimated neuromotor state with a reference state, yielding residual acoustic energy and latent displacement measures that can be interpreted and audited. The framework specifies training losses, invariance constraints, uncertainty handling, decision thresholds, and evaluation endpoints suitable for early detection studies. It also identifies failure modes that can invalidate apparent biomarker performance, including disentanglement leakage, microphone shift, label noise, and over-normalization of disease-related variation. The approach provides a technically coherent route from generative speech representations to reproducible, clinically meaningful acoustic biomarkers without treating synthetic plausibility as diagnostic evidence.

Keywords

acoustic biomarkers; self-supervised speech representation; generative latent modeling; counterfactual normalization; neurological disorder detection; calibration

1. Introduction

Neurological disorders commonly affect speech production and language generation before they affect functional independence. In Parkinson's disease, hypokinetic dysarthria arises from basal ganglia dysfunction and manifests as reduced prosodic variation, imprecise consonants, altered voice quality, and variable speech timing [1,2]. In Alzheimer's disease and related dementias, narrative speech can show word-finding difficulty, reduced semantic content, simplified syntax, and altered pause structure [3,4]. These clinical observations motivate the use of speech as a low-burden, repeatable, and remotely collectable signal for early detection. The central problem is not whether speech differs statistically between groups; it is whether a speech-derived measurement can be defined precisely enough to support prospective screening, longitudinal monitoring, and clinical decision-making.

Existing acoustic analysis offers useful measurement principles but limited transportability. Quantitative studies have characterized vowel space, formant dynamics, jitter, shimmer, harmonic-to-noise ratio, pause duration, and speech rate in early untreated Parkinson's disease [5]. Machine-learning studies using sustained phonation and speech tasks have demonstrated that handcrafted acoustic descriptors can classify Parkinsonian speech with high accuracy in controlled cohorts [6]. Such features are attractive because they map to physiological mechanisms and can be computed transparently. Their weakness is sensitivity to segmentation, microphone response, background noise, speaking task, language, and speaker anatomy. A feature that discriminates cases from controls in one recording setting may encode site-specific rather than disease-specific variation.

Self-supervised speech models provide a different starting point. Models such as wav2vec 2.0 and HuBERT learn contextual representations by predicting masked or quantized acoustic units from large unlabeled corpora [7,8]. Their layers encode information ranging from low-level spectral structure to phonetic and lexical content, and they reduce dependence on handcrafted feature engineering. The difficulty is that a high-dimensional representation is not yet a biomarker. Downstream classifiers trained on such representations can perform well while remaining opaque, vulnerable to shortcut learning, and poorly calibrated across populations. Generative speech models add the possibility of controlled synthesis and counterfactual analysis, but they introduce identifiability problems: a model may reproduce plausible speech while failing to isolate the physiological factor relevant to disease.

This paper develops a framework for acoustic biomarker extraction using generative speech representations. The contribution is not a single classifier but a measurement architecture: a self-supervised encoder produces segmental speech representations; a conditional generative model decomposes those representations into content, speaker style, and a latent neuromotor state; biomarkers are computed as deviations from a normative manifold and as residual acoustic energy after counterfactual resynthesis. The framework is designed to make the biomarker definition explicit, to separate disease-related variation from nuisance variation, and to subject the resulting scores to calibration, invariance, and decision-curve evaluation. The argument proceeds from clinical and acoustic foundations to the generative representation model, then to evaluation logic and technical boundary conditions.

2. Clinical and Acoustic Foundations

The clinical value of an acoustic biomarker depends on the disorder mechanism and the intended detection task. Parkinsonian speech reflects reduced amplitude and speed of articulatory movements, impaired laryngeal control, and altered respiratory support. These mechanisms can produce measurable changes in fundamental frequency range, intensity, vowel formants, voice onset and offset, consonant closure, and pause distribution. A sensitive biomarker might therefore target dynamic articulatory transitions rather than static sustained vowels alone. Alzheimer's disease and related dementias involve lexical retrieval and discourse planning, so informative measures include pause location, repetition, pronoun density, semantic coherence, syntactic complexity, and speech rate during picture description. These are not purely acoustic constructs; they integrate acoustic timing with linguistic content. A generative representation can encode both sources, but the biomarker definition must specify which variation is considered pathological and which is attributable to task, language, or speaker style.

Early detection imposes stricter requirements than retrospective case-control classification. The target population is often symptomatic but undiagnosed, or at risk because of family history, subjective cognitive decline, rapid eye movement sleep behavior disorder, or mild motor signs. Prevalence is lower than in a specialist clinic, so positive predictive value is sensitive to specificity and calibration. A speech score that ranks cases above controls may still be unsuitable for triage if its risk estimates are miscalibrated at the operating threshold. The relevant endpoints include discrimination, calibration slope and intercept, decision-curve net benefit, longitudinal change, and time to clinical diagnosis. A useful biomarker should also be stable in unchanged individuals, sensitive to meaningful progression, and reproducible across recording devices and languages.

The acoustic foundation must therefore distinguish three quantities: the observed signal, the physiological state that generates it, and the clinical label. Speech is produced by a chain involving respiratory pressure, laryngeal vibration, vocal tract filtering, articulatory timing, and central linguistic planning. The observed waveform is also shaped by microphone frequency response, room reverberation, compression, and sampling rate. Generative modeling is attractive because it can represent these factors explicitly, but it cannot eliminate them by declaration. The model must be constrained so that content and speaker style are preserved during counterfactual transformation, and the resulting biomarker must be tested for invariance to nuisance factors without erasing disease-related variation correlated with age, sex, language, or medication status.

3. Methodological Framework

3.1 Problem formulation

Let x denote a speech waveform or a set of task-specific recordings from one participant, a denote observed covariates such as age, sex, language, recording device, and medication timing, and y denote the clinical endpoint. For early detection, y may be a clinician-confirmed diagnosis at follow-up, a consensus label from a multidisciplinary assessment, or a longitudinal progression indicator. The framework learns an encoder f_θ that maps the waveform into a sequence of latent vectors. The output is not a diagnosis but a set of segment-level biomarker scores b(x), a calibrated risk estimate, and an uncertainty summary. The biomarker function must be defined on clinically meaningful speech segments, such as

sustained vowels, diadochokinetic repetitions, reading passages, and picture-description narratives.

The encoder is written as follows, with $x_{1:n}$ denoting the input sequence and $z_{1:t}$ the resulting representation sequence.

$$z_{1:t} = f\theta(x_{1:n})$$

Here θ denotes encoder parameters, N is the number of input samples, and T is the number of output frames or segments. The mapping may be obtained from a pretrained self-supervised model and adapted with task-specific fine-tuning. Frame-level representations are pooled over segments W_i chosen by forced alignment or robust segmentation. The segment definition is part of the biomarker specification because disease effects may be concentrated in vowels, consonant transitions, pauses, or narrative boundaries.

3.2 Self-supervised representation and pooling

Self-supervised speech encoders are appropriate because they learn acoustic and phonetic structure without requiring large disease-labeled corpora. wav2vec 2.0 learns discrete latent units through contrastive prediction and masked modeling, while HuBERT uses clustered hidden units as prediction targets [7,8]. Their intermediate layers contain complementary information: early layers retain spectral and voice-quality information, whereas later layers emphasize phonetic and lexical content. A biomarker framework can exploit this hierarchy by learning layer weights rather than selecting one layer by convention. The pooling operation must also be specified. Mean pooling over a vowel may capture stable voice quality, while statistics of frame-to-frame change may capture articulatory instability. For narrative speech, pause-aware pooling can separate within-word timing from between-utterance planning.

The generative component operates on these representations rather than directly on raw audio. This choice reduces the dimensionality of the generative problem and allows the model to use pretrained acoustic knowledge. It also creates a risk: if the encoder has already discarded a disease-relevant cue, no downstream generative model can recover it. The framework therefore requires layer-wise sensitivity analysis and a comparison against handcrafted features and raw-waveform models. The encoder should be treated as a measurement device whose invariance properties are tested, not assumed.

3.3 Generative latent decomposition

The generative model assumes that each representation can be decomposed into content c_t, speaker style s, and a latent neuromotor state m_t. Content represents what is said, speaker style represents stable vocal-tract and habitual characteristics, and neuromotor state represents time-varying control relevant to disease. The decomposition is expressed as follows, with additive residual noise.

$$z_t = G(c_t, s, m_t) + \varepsilon_t,\ \varepsilon_t \sim N(0, \Sigma_t)$$

G is a generative mapping, ε_t is residual variation, and Σ_t is a conditional covariance matrix. The covariance may depend on segment type and recording quality; it is not assumed constant across speakers or tasks. The latent variables are not observed, so the model must be trained with reconstruction, distribution-matching, and weak supervision. Variational autoencoders provide a probabilistic objective for latent-variable learning [9], generative adversarial networks provide a flexible adversarial objective for realistic synthesis [10], and diffusion models provide a denoising-based generative process that can model complex

conditional distributions [11]. The framework is compatible with any of these families, but the choice affects identifiability, sampling cost, and the reliability of counterfactual resynthesis.

Disentanglement requires more than an architectural split. Without constraints, the model can place speaker information in m_t or disease information in s. A practical strategy combines three constraints. First, content consistency is enforced by requiring that resynthesis with changed m preserves phonetic identity under an automatic speech recognition or phonetic discriminator. Second, speaker consistency is enforced by requiring that voice identity remains stable when m is changed. Third, disease-related supervision is applied only to the m pathway, with gradient isolation or adversarial removal of the clinical label from the content and style pathways. Style-transfer architectures for singing-voice synthesis demonstrate how content and style can be computationally separated for controlled resynthesis [12], but clinical speech requires explicit invariance testing because the relevant neuromotor factor is not equivalent to artistic style. A model that generates convincing speech can still entangle pathology with recording quality or speaker identity.

3.4 Biomarker extraction and counterfactual normalization

The biomarker is defined as a deviation from a normative neuromotor state m_0, conditional on content and speaker style. For a segment W_i, let δ_t denote the residual between the observed representation and the representation generated under m_0. The segment-level biomarker is the Mahalanobis energy of this residual.

$$b_i(x) = (1/|W_i|) \sum_t \in W_i \, \delta_t{}^T \Sigma_t{}^{-1} \, \delta_t, \; \delta_t = z_t - G(c_t, s, m_0)$$

The state m_0 represents a reference neuromotor condition, not a healthy person in general. It must be conditioned on age, sex, language, and task where appropriate. The covariance Σ_t weights residual directions according to their expected variability under the reference model. A large $b_i(x)$ indicates that the observed segment lies far from the normative manifold in directions that the model attributes to neuromotor variation. This definition differs from a classifier score because it is anchored to a generative counterfactual and can be decomposed by segment type, acoustic scale, and latent direction.

The corresponding acoustic residual is obtained by decoding the counterfactual representation and subtracting the resynthesized segment from the observed segment. Let D denote the decoder and $x(W_i)$ the observed segment.

$$r_i(x) = x(W_i) - D\big(G(c(W_i), s, m_0)\big)$$

The residual $r_i(x)$ provides a time-domain or spectral check on the latent biomarker. If $b_i(x)$ is large but $r_i(x)$ is dominated by high-frequency noise, the latent score may be reflecting channel characteristics rather than neuromotor control. If both the latent energy and an interpretable acoustic residual increase in the expected frequency bands, the evidence is more coherent. The residual should be analyzed with physically meaningful summaries such as formant-transition slope, harmonic-to-noise ratio, pause duration, and spectral flux, not only with aggregate energy.

3.5 Training objective and constraints

The generative model is trained with a composite objective that balances reconstruction, latent regularization, adversarial realism, cycle consistency, and supervised prediction of clinical labels or severity scores. A general form is given below.

$$L(\theta, \varphi) = L_r ec + \lambda_1 L_K L + \lambda_2 L_a dv + \lambda_3 L_c yc + \lambda_4 L_s up$$

θ denotes encoder parameters, φ denotes generative and decoder parameters, L_rec is reconstruction loss, L_KL regularizes the latent posterior toward its prior, L_adv encourages realistic resynthesis, L_cyc enforces consistency after content-style-neuromotor transformations, and L_sup is a supervised loss on available clinical labels or severity scales. The coefficients λ_1 through λ_4 are not universal constants; they express trade-offs between realism, identifiability, and predictive signal. A high adversarial weight may improve perceptual quality while destabilizing latent inference. A high supervised weight may improve classification while encouraging leakage of the label into content or style. The framework therefore treats these weights as design variables to be selected by nested validation and sensitivity analysis.

Waveform-level generative models such as autoregressive and diffusion-based synthesizers can improve acoustic realism [11,13], but realism is not the target. The target is a counterfactual that changes the neuromotor state while preserving content, speaker identity, and recording conditions. A perceptually convincing synthesis can still be clinically invalid if it changes the wrong factor. The evaluation of the generative component should therefore include factor-specific interventions: changing m while holding c and s fixed, changing s while holding c and m fixed, and changing recording channel while holding all other factors fixed. Each intervention tests a different invariance property required by the biomarker definition.

3.6 Risk estimation, uncertainty, and longitudinal scoring

A biomarker score becomes clinically useful only after it is mapped to a calibrated risk. Let hψ be a risk function that combines the biomarker vector b(x), latent representation summaries, and covariates a. The predictive distribution integrates over the inferred latent state.

$$P(y = 1 \mid x) = \int \sigma(h\psi(z, b(x), a))\, p\varphi(z \mid x)\, dz$$

σ is a link function producing a probability, and pφ(z|x) is the posterior over latent representations. The integral can be approximated by Monte Carlo sampling during inference. This formulation separates aleatoric uncertainty, represented by the conditional distribution of z, from epistemic uncertainty, which should be estimated through ensemble or Bayesian approximations. Calibration must be assessed at the intended operating threshold because a model can have good discrimination and poor calibration [15]. Decision-curve analysis is appropriate for comparing net benefit across threshold probabilities and for making the clinical trade-off explicit [16].

Longitudinal scoring requires a different statistic from cross-sectional deviation. For a participant with repeated recordings at times t_1 and t_2, a progression score can be defined as the rate of change in segment-specific biomarker energy.

$$\Delta_i = \frac{\left[b_i\big(x(t_2)\big) - b_i\big(x(t_1)\big)\right]}{(t_2 - t_1)}$$

Δ_i is interpretable only if the speech task, recording condition, and medication state are comparable across visits. A change in microphone or language can produce a spurious slope. Longitudinal models should therefore include visit-level nuisance covariates, measurement-error models, and tests of within-person reliability. A biomarker that is sensitive to disease progression but unstable in unchanged individuals is not suitable for monitoring.

4. Methodological Evaluation and Error Analysis

4.1 Study design and endpoints

A credible evaluation of the framework requires a prospective or rigorously designed longitudinal cohort with clinician-confirmed outcomes, not only retrospective convenience samples. Participants should complete standardized tasks that stress the relevant mechanisms: sustained vowel phonation for voice stability, diadochokinetic repetition for articulatory timing, reading for segmental accuracy, and picture description for spontaneous language. The same task battery should be collected across sites, with device metadata and acoustic calibration recordings. The primary endpoint for early detection should be diagnosis at a prespecified follow-up, not diagnosis at the same visit, when the claim concerns early detection. Secondary endpoints should include clinical severity, conversion status, and longitudinal change.

The analysis plan should prevent speaker leakage by splitting participants rather than recordings. Hyperparameter selection should occur inside nested cross-validation, and external validation should use a separate site or population. Discrimination should be reported with confidence intervals and compared against handcrafted acoustic features, raw waveform classifiers, and x-vector speaker embeddings [14]. Calibration should be assessed with reliability plots, calibration slope, and intercept [15]. Clinical utility should be assessed with decision-curve analysis [16]. Subgroup analyses should examine age, sex, language, education, medication state, and recording device. These analyses are not optional refinements; they determine whether the biomarker is measuring disease or measuring access to a particular clinical environment.

4.2 Confounding, invariance, and domain shift

Speech biomarkers are exposed to nuisance variation from microphone frequency response, room acoustics, compression, speaking effort, fatigue, language, dialect, and comorbidity. Domain adaptation theory shows that predictive performance depends on both the divergence between source and target distributions and the labeling function mismatch [17]. A model trained at one site may therefore fail at another even when the disease mechanism is unchanged. The generative framework can mitigate this problem by treating recording channel and speaking style as explicit nuisance factors, but the mitigation must be measured. One useful regularizer penalizes systematic dependence of the biomarker on nuisance strata a.

$$L_inv(\varphi) \;=\; \sum_a \,||\, E[b(x) \mid a] \;-\; E[b(x)] \,||^2$$

a indexes nuisance or covariate strata, and E[b(x)|a] is the conditional mean biomarker within a stratum. The penalty encourages the biomarker to have similar average behavior across strata. It must be applied selectively because some covariates, such as age, are legitimately correlated with disease risk. Over-applying invariance can remove true pathological signal. Invariance should therefore be evaluated on nuisance factors that are not causal for the disease mechanism, such as microphone model or room reverberation, and reported separately from invariance to demographic factors.

4.3 Failure modes and sensitivity

The framework has several failure modes that can create false confidence. Disentanglement leakage occurs when the neuromotor latent absorbs speaker or channel information; it can produce large biomarker scores for loud speakers or for recordings made

with a particular device. Over-normalization occurs when the counterfactual removes disease-related variation along with nuisance variation; it can reduce sensitivity in exactly the early stages the biomarker is intended to detect. Generative hallucination occurs when the decoder synthesizes plausible content that is not supported by the observed acoustics; residual analysis can detect this if the residual is dominated by high-frequency or unstructured noise. Label noise and delayed diagnosis can weaken supervision and bias risk estimates. Selection bias can make a model appear effective because cases were referred from a specialized clinic with distinctive recording conditions.

Sensitivity analysis should test the biomarker against alternative encoder layers, pooling windows, covariance estimators, and generative families. It should quantify the effect of segment choice, alignment error, and speaking task. It should also test whether the biomarker remains associated with the clinical endpoint after adjustment for covariates and whether the association survives external validation. A biomarker that changes substantially under reasonable analytical choices is not ready for clinical interpretation. Conversely, a stable biomarker with a clear residual interpretation can be investigated mechanistically even before it is used for diagnosis.

4.4 Comparison with alternative approaches

Handcrafted acoustic features remain the appropriate baseline because they are interpretable and computationally transparent [5,6]. Their limitations are well known, but they provide a check against which generative representations must justify their added complexity. Self-supervised classifiers offer stronger representation learning but can obscure the measurement definition. Generative models offer counterfactual control and data augmentation, but they introduce identifiability and synthesis-validity risks. A hybrid design is therefore most defensible: generative representations supply the latent space and counterfactual machinery, while physically interpretable acoustic residuals supply a verification layer. The combination allows the framework to test whether a latent deviation corresponds to a measurable change in voice, articulation, or timing.

5. Technical Discussion

5.1 Data efficiency, synthesis, and privacy

Large self-supervised models reduce the need for disease-labeled speech, but they do not eliminate the need for clinically characterized cohorts. Generative models can create controlled perturbations for sensitivity testing and can augment underrepresented recording conditions. Synthetic speech can also support privacy-preserving sharing if residual risk is carefully managed. However, synthetic data cannot substitute for external validation because the generative model may reproduce the biases of its training corpus. Augmentation is useful when it expands nuisance variation without changing the clinical endpoint. It is harmful when it creates unrealistic combinations of disease, speaker, and recording factors that teach the classifier spurious associations.

5.2 Clinical translation and decision thresholds

The clinical use of an acoustic biomarker depends on intended use, prevalence, and the consequence of false positives and false negatives. A screening tool for a low-prevalence condition requires high specificity and calibrated risk estimates; a monitoring tool for progression requires sensitivity to within-person change. The risk function should be reported with a prespecified decision threshold and a decision-curve analysis rather than as a single

area-under-the-curve value [16]. Biomarker validation principles from neuroimaging and digital medicine emphasize the need for a clear clinical question, a defined target population, and independent validation before claims of clinical utility [19,20]. Speech adds specific challenges because the measurement is remote, device-dependent, and culturally variable. These challenges make standardization and transparency central rather than peripheral.

5.3 Boundary conditions

The framework does not establish that a generative latent variable corresponds to a physiological mechanism. It establishes a measurement model whose validity can be tested. The normative state m_0 is conditional and context-dependent; it is not a universal healthy template. The biomarker is segment-specific; a score averaged over all speech can obscure focal deficits. The generative counterfactual is model-dependent; if the decoder changes content or speaker identity, the residual is not a clean measure of neuromotor deviation. The framework is therefore best understood as a disciplined way to construct and audit acoustic biomarkers, not as a replacement for clinical diagnosis. Its scientific value lies in making assumptions explicit enough to be falsified.

6. Conclusion

This manuscript has developed a methodological framework for acoustic biomarker extraction from generative speech representations, motivated by the need for early detection of neurological disorders and by the fragility of conventional acoustic features. The framework couples self-supervised speech encoding, conditional latent decomposition, counterfactual resynthesis, segment-level deviation scoring, calibrated risk estimation, and longitudinal change measurement. Its central technical claim is that a generative representation becomes a biomarker only when the measurement definition, the counterfactual reference state, the nuisance-invariance constraints, and the clinical decision rule are specified together. The biomarker is not the classifier output; it is a decompositional measurement that can be checked against acoustic residuals and tested for stability across devices, speakers, languages, and sites.

The framework also clarifies why plausible synthesis is insufficient evidence of clinical validity. Disentanglement can leak, normalization can erase disease signal, and domain shift can create site-specific scores. These risks are not incidental; they are central design problems. A defensible research program should compare generative biomarkers with handcrafted and self-supervised baselines, report calibration and decision curves, test invariance to noncausal nuisance factors, and validate in independent cohorts with delayed clinical outcomes. The contribution of this work is to make those requirements explicit and to provide a coherent analytical structure for extracting interpretable acoustic biomarkers from generative speech representations.

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. 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

3. Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. Journal of Alzheimer's Disease, 49(2), 407–422. https://doi.org/10.3233/JAD-150520

4. Ahmed, S., Haigh, A.-M., de Jager, C. A., & Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. Brain, 136(12), 3727–3737. https://doi.org/10.1093/brain/awt269

5. Rusz, J., Cmejla, R., Ruzicka, E., & Ruszova, H. (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

6. Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J., & Ramig, L. O. (2012). Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Transactions on Biomedical Engineering, 59(5), 1264–1271. https://doi.org/10.1109/TBME.2012.2183367

7. 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).

8. 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

9. Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In Proceedings of the 2nd International Conference on Learning Representations.

10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems (Vol. 27, pp. 2672–2680).

11. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems (Vol. 33, pp. 6840–6851).

12. 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).

13. van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.

14. Jeancolas, L., Petrovska-Delacrétaz, D., Mangone, G., Benkelfat, B.-E., Corvol, J.-C., Vidailhet, M., Roze, E., & Lehéricy, S. (2021). X-vectors: New quantitative biomarkers for early Parkinson's disease detection from speech. Frontiers in Digital Health, 3, 578369. https://doi.org/10.3389/fdgth.2021.578369

15. 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).

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. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. Machine Learning, 79(1–2), 151–175. https://doi.org/10.1007/s10994-009-5152-4

18. Marquand, A. F., Rezek, I., Buitelaar, J., & Beckmann, C. F. (2016). Understanding heterogeneity in clinical cohorts using normative models: beyond case-control studies. Biological Psychiatry, 80(7), 552–561. https://doi.org/10.1016/j.biopsych.2015.12.023

19. Woo, C.-W., Chang, L. J., Lindquist, M. A., & Wager, T. D. (2017). Building better biomarkers: brain models in translational neuroimaging. Nature Neuroscience, 20(3), 365–377. https://doi.org/10.1038/nn.4478

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