SpeechDx: A Multi-Task Benchmark for Clinical Speech AI

Sejal Bhalla, Larry Kieu, Aina Merchant, Eyal de Lara, Alex Mariakakis

University of Toronto, Canada
sejal@cs.toronto.edu

Abstract

Speech offers a uniquely informative window into health by simultaneously engaging neurological, motor, respiratory, and vocal systems. Current clinical speech AI methods have largely progressed through isolated condition-specific studies, making results difficult to compare and generalization difficult to assess. We introduce SpeechDx, a large-scale benchmark for clinical speech AI spanning 12 datasets and 27 tasks across diverse health conditions. To enable evaluation across shared clinical mechanisms, SpeechDx structures tasks by the stage of speech production they disrupt: conceptualization, formulation, and articulation. The benchmark tests generalization by including tasks with limited labeled data and evaluating the same health condition across multiple datasets, distinguishing clinically meaningful patterns from dataset artefacts. We systematically evaluate 12 state-of-the-art audio encoders across all tasks and under zero-shot cross-condition transfer. Results show that large-scale speech models represent the strongest overall baselines, domain-specific models improve performance only on closely matched tasks, and no current representation generalizes reliably across the clinical speech landscape. SpeechDx establishes a shared evaluation framework for tracking progress toward general-purpose clinical speech representations.

1 Introduction

Speech has been widely studied as a digital biomarker of human health. Its production requires the coordinated action of the respiratory, vocal, neurological, and cognitive systems, so disruptions to any of these systems due to disease or dysfunction leave measurable traces in the voice. Unlike many clinical assessment tools, speech can be captured non-invasively, remotely, and at negligible cost, making it particularly well-suited for continuous monitoring, population-level screening, and longitudinal disease tracking. Over the past few decades, this has motivated a substantial body of research exploring the automation of speech analysis for a wide range of conditions such as COVID-19 [1, 2], dysarthria [3, 4, 5], Parkinson's disease [6, 7, 8], Alzheimer's disease [9, 10, 11], depression [12, 13, 14], and vocal pathologies [15, 16]. Dedicated challenges further demonstrate the potential of deep learning for health assessment from speech [17, 18, 19, 20, 21, 22].

Despite this potential, very few systems have achieved real-world clinical deployment. A recurring obstacle is the fragmented nature of the field. Research has progressed in silos, with models trained and evaluated on individual datasets under inconsistent protocols, limiting generalizability across conditions. Even within a single condition, models trained on small controlled corpora consistently fail to generalize to unseen data. This failure is largely attributed to models learning spurious correlations from confounding factors in the data — differences in recording conditions, demographic composition, and acquisition hardware — rather than the underlying clinical signal [23, 24, 25]. The field lacks a standardized evaluation framework for quantifying progress, assessing generalization across datasets and conditions, and identifying robust modeling approaches.

Preprint.

We address this gap by introducing SpeechDx, a clinical speech AI benchmark comprising **12 publicly available** speech datasets with **27 tasks** spanning **nine health and affective conditions**. One of SpeechDx's distinguishing features is its organization around the speech production process. Following the framework proposed by Berisha and Liss [26], we classify conditions and tasks according to the stage of speech production at which the condition exerts its primary effect: (1) conceptualization, when a communicative intention is formed; (2) formulation, when that intention is encoded into linguistic and phonological structure; and (3) articulation, where respiratory and motor systems execute the spoken output. Diverse health conditions disrupt different stages of this cascade, producing distinct types of acoustic irregularity.

We use SpeechDx to evaluate 12 state-of-the-art audio and speech encoders representing different representation learning paradigms. We first benchmark the models across all tasks, establishing how well the pretrained representations encode clinically relevant information. Second, we conduct a zero-shot cross-condition analysis, revealing shared acoustic structure across conditions and identifying where generalization breaks down. To summarize, our contributions are as follows:

- We introduce SpeechDx, the first large-scale benchmark specifically designed to advance clinical speech AI. This benchmark comprises 12 datasets and 27 tasks across nine health conditions. The codebase is available at `https://anonymous.4open.science/r/SpeechDx-F584`.
- We provide a systematic evaluation of 12 state-of-the-art audio encoders, establishing a standardized and reproducible baseline for performance comparison across the clinical domains.
- We conduct a zero-shot cross-condition transfer analysis to reveal which conditions share learnable acoustic structure and where transfer degrades.

2 Background and Related Work

2.1 Speech as a Biomarker

Deviations in vocal attributes such as pitch, intensity, resonance, and temporal structure constitute measurable markers of disordered voice production. Such deviations arise from perturbations in the biological systems underlying speech generation [27, 28]. This sensitivity is fundamental to the utility of speech as a biomarker [29] and contributes to its recognition as a vital sign [30]. Its non-invasive and low-cost acquisition further enables repeated and longitudinal measurements that support applications in screening, monitoring, and prognosis.

Berisha and Liss [26] propose a framework that organizes health conditions by how directly they disrupt the speech production mechanism. At one end of the spectrum, conditions directly impact the mechanics of acoustic speech production. On the other end, conditions disrupt cognitive-affective processes or linguistic planning, resulting in changes to speech content and formulation. SpeechDx adopts this framework as the basis for task selection and organization.

2.2 Clinical Speech AI

The field of clinical speech AI broadly studies computational methods that extract acoustic, prosodic, and linguistic information from speech to infer clinically relevant outcomes [31]. Early work in the field relied on hand-engineered features such as MFCCs, jitter, shimmer, and prosodic descriptors [32]. These representations have proven to be useful for tasks spanning COVID-19 detection [33], Parkinson's disease classification [34], and depression screening [35], among others [36, 37, 38]. More recent work has shifted toward deep neural networks trained directly on audio waveforms or spectrograms [39, 40, 41, 42, 43, 44]. However, both conventional and deep approaches are typically optimized for individual conditions. Although these models often achieve strong within-dataset performance, they frequently fail under distribution shift [23, 31, 45].

Innovations in self-supervised learning across vision [46, 47, 48] and language [49, 50, 51] have shown that it is possible to learn universal representations that transfer across several downstream tasks with minimal adaptation. An analogous capability is desirable in clinical speech AI since labeled datasets are small and costly to collect. A disease-agnostic representation could provide a shared basis for learning across multiple clinical tasks, reducing dependence on large labeled datasets for each disease. Initial studies using representations from models such as wav2vec 2.0 [52], HuBERT [53], and WavLM [54] have shown promising results on clinical speech tasks [55, 56, 57, 58, 59, 60], suggesting that pretrained speech encoders capture information relevant to health assessment.

Recent work has therefore shifted toward health-oriented audio representations. WavRx [61] extended WavLM with a modulation dynamics module to capture respiration and articulation abnormalities, while HeAR [62] included a masked autoencoder trained on 313 million health acoustic clips. Both represent important steps towards general audio representations for health. However, their evaluation scope remains limited: WavRx covers six datasets and four pathologies, while HeAR was tested on tasks spanning respiratory sounds like coughing and breathing. The broader scope of speech-affecting diseases spanning motor, cognitive-linguistic, and affective conditions remains largely unaddressed.

Standardized benchmarks have driven systematic progress in adjacent domains; SUPERB [63] established a shared multi-task evaluation protocol that accelerated semantic speech research, while HEAR [64] provided a framework that advanced general audio representation learning. SpeechDx serves this purpose for clinical speech AI, providing a catalyst for the development of general-purpose health audio representations that transfer across clinical tasks and populations.

3 Datasets and Tasks

Subject to clinical relevance and data access constraints, the dataset curation process yielded 12 datasets with 27 downstream tasks. Table 1 summarizes the key characteristics of these datasets and tasks, but more detailed descriptions can be found in Appendices A and B respectively. Below, we describe how the datasets are positioned under Berisha and Liss' framework for speech production [26].

3.1 Conceptualization Disorders

Conceptualization is the first stage of speech production, where the speaker forms a communicative intention and decides what to express. This stage is driven by cognitive and affective processes — how the speaker feels, what they want to communicate, and how much they elaborate. Disruptions here alter cognitive drivers that result in subtle acoustic changes such as reduced speaking rate, flattened pitch variation, truncated responses, and altered patterns of emphasis and timing.

Depression and emotional dysregulation are among the most common conditions affecting this stage. Depression reduces psychomotor drive and cognitive engagement, while emotional state more broadly shapes the prosodic and temporal properties of spoken output. Both are clinically significant conditions for which scalable, low-cost assessment tools are actively needed. We use EDAIC-WOZ [65] to probe the presence and severity of depression, and RAVDESS [66] and IEMOCAP [67] to probe multi-class emotion recognition in acted and naturalistic speech, respectively.

3.2 Formulation Disorders

The formulation stage involves encoding communicative intent into linguistic structures through word selection, syntactical rules, and phonetic sequencing. Disrupted formulation often results in speech that is acoustically fluent yet contains unusual sentence structures or phonemic substitutions.

Language disorders following neurological damage and neurodegenerative disease are the most common disruptors of this stage. Aphasia, arising from stroke or focal brain injury, directly impairs lexical retrieval and syntactic encoding. On the other hand, Alzheimer's disease progressively disrupts semantic memory and word access, with downstream effects on the coherence of spoken output. These conditions represent two of the most clinically prevalent acquired communication disorders in adults, and detecting them from speech alone is highly valuable for tracking disease progression. We use standardized discourse recordings from AphasiaBank [69] to detect the presence of aphasia and DementiaBank [68] to detect the presence and severity of Alzheimer's disease.

3.3 Articulation Disorders

Articulation is the stage at which planned speech is executed through coordinated movement of the respiratory, laryngeal, and articulatory systems. This execution can be broadly decomposed into two subsystems: (1) the neuromuscular articulatory subsystem, which modulates airflow into intelligible speech, and (2) the phonatory and respiratory subsystem, which generates airflow in the first place.

Table 1: The datasets and tasks comprising SpeechDx. Task types are either classification (C), multi-label classification (M), or regression (R). Dataset splits either entail a single train-validation-test split (TVT) or 5-fold subject-wise cross-validation (5-fold). Sample sizes denote the number of recordings, with the number of unique subjects shown in parentheses.

Category	Dataset	ID	Task	Type	Split	Samples (Subjects)
Conceptualization	EDAIC-WOZ [65]	T1	Depression / healthy	C	TVT	163 (163) / 56 (56) / 56 (56)
		T2	PHQ-8 score	R	TVT	163 (163) / 56 (56) / 56 (56)
	RAVDESS [66]	T3	Emotion classification	C	5-fold	1,140 (19) / 300 (5)
		T4	Negative / non-negative emotion	C	5-fold	1,140 (19) / 300 (5)
	IEMOCAP [67]	T5	Emotion classification	C	5-fold	5,843 (8) / 1,537 (2)
		T6	Negative / non-negative emotion	C	5-fold	5,843 (8) / 1,537 (2)
Formulation	DementiaBank [68]	T7	Dementia / healthy	C	TVT	237 (237) / 8 (8) / 46 (46)
		T8	MMSE score	R	TVT	236 (236) / 8 (8) / 46 (46)
	AphasiaBank [69]	T9	Aphasia / healthy	C	TVT	740 (143) / 160 (21) / 237 (42)
Articulation (Neuromuscular)	TORGO [70]	T10	Dysarthria / healthy	C	5-fold	7,605 (12) / 1,811 (3)
		T11	Dysarthria severity	R	5-fold	2,379 (6) / 802 (2)
	UASpeech [71]	T12	Dysarthria / healthy	C	5-fold	16,830 (22) / 4,590 (6)
	MDVR-KCL [72]	T13	Parkinson's / healthy	C	5-fold	59 (30) / 14 (7)
		T14	UPDRS-5 score	R	5-fold	59 (30) / 14 (7)
		T15	UPDRS-18 score	R	5-fold	59 (30) / 14 (7)
		T16	H&Y score	R	5-fold	59 (30) / 14 (7)
	KSoF-C [73]	T17	Disfluency / healthy	C	5-fold	4,307 (31) / 1,036 (6)
		T18	Disfluency classification	M	5-fold	4,394 (30) / 1,203 (7)
Articulation (Phonatory / Respiratory)	COVID-19 Sounds [74]	T19	Symptomatic / healthy (official subset)	C	TVT	6,565 (4,635) / 943 (663) / 1,948 (1,325)
		T20	Symptomatic / healthy	C	TVT	37,353 (25,252) / 5,271 (3,608) / 10,761 (7,216)
		T21	COVID-19 / non-COVID-19 (official subset)	C	TVT	1,017 (700) / 141 (100) / 324 (200)
		T22	COVID-19 / non-COVID-19	C	TVT	5,606 (3,711) / 804 (531) / 1,609 (1,061)
		T23	Symptom classification	M	TVT	19,541 (15,459) / 2,914 (2,209) / 5,591 (4,418)
	Coswara [75]	T24	Symptomatic / healthy	C	TVT	3,786 (1,896) / 542 (271) / 1,083 (542)
		T25	COVID-19 / non-COVID-19	C	TVT	2,915 (1,459) / 418 (209) / 834 (418)
		T26	Symptom classification	M	TVT	1,459 (731) / 210 (105) / 417 (209)
	AVFAD [76]	T27	Vocal pathology / healthy	C	TVT	3,968 (496) / 568 (71) / 1,135 (142)

Neuromuscular Disorders. These disorders reduce the precision, speed, and coordination of the vocal articulators. This reduction results in speech that is imprecise and less intelligible. We represent this class of disorders with two examples: (1) dysarthria, a group of motor speech disorders characterized by weakness or incoordination of the speech musculature; and (2) disfluency, which disrupts the timing and sequencing of speech production. TORGO [70] and UASpeech [71] support dysarthria detection and severity estimation among speakers with diverse etiology. MDVR-KCL [72] targets hypokinetic dysarthria as observed in Parkinson's disease, which is characterized by reduced vocal loudness, monotone pitch, and imprecise articulation. Finally, KSoF-C [73] provides fluency-labeled recordings to enable disfluency classification and stuttering detection.

Phonatory and Respiratory Disorders. The respiratory and phonatory subsystem is responsible for the airflow and vocal fold vibration that underlie all voiced sounds. Conditions affecting this subsystem disrupt speech at its source, producing what could theoretically be the most directly measurable acoustic deviations. We consider both disease detection and symptom characterization in this category of disorders. The latter is particularly relevant for respiratory conditions that are accompanied by cough, cold, and breathing difficulty, each of which exhibits distinct acoustic patterns during speech. We use COVID-19 Sounds [74] and Coswara [75] for respiratory symptom characterization and COVID-19 detection, and AVFAD [76] for vocal pathology detection.

4 Methods

4.1 Audio Preprocessing

All audio is resampled to 16 kHz, mono-channeled, and normalized prior to modeling. To accommodate variable-length recordings, inputs shorter than a model's minimum accepted length are zero-padded, and those exceeding the maximum accepted length are divided into non-overlapping chunks whose embeddings are mean-pooled to produce a single representation.

4.2 Models

We evaluate 12 state-of-the-art audio and speech models that have been widely used for spoken language, paralinguistic, and health-related tasks. These models are selected to cover a range of training data sources (general speech, general audio, and domain-specific audio), supervision paradigms, and learning objectives, enabling comparisons across varying levels of data scale and domain alignment. Model details are provided in Table 4, with expanded descriptions in Appendix C.

Speech Models. Following prior work demonstrating the use of self-supervised speech representations in clinical and paralinguistic applications [55, 58, 60, 77, 78, 57], we evaluate wav2vec 2.0 [52], HuBERT [53], WavLM [54], and MMS [79]. These models utilize large-scale self-supervised speech pretraining followed by supervised fine-tuning for automatic speech recognition, but they differ in their pretraining objectives and data. wav2vec 2.0 uses contrastive learning over quantized representations, HuBERT uses masked prediction with offline cluster targets, WavLM adds a denoising objective, and MMS extends this framework to over 1,400 languages. We additionally include the Qwen3-TTS-Tokenizer [80], a neural audio codec trained on multilingual speech corpora with a self-supervised reconstruction objective; and Whisper [81], a fully supervised model trained on weakly labeled speech via sequence-to-sequence learning.

General Audio Models. General audio models are pretrained on broad audio corpora such as AudioSet [82], which include but are not limited to speech. This exposure provides them with a wider acoustic distribution than that of speech-only models. We include AudioMAE [83] and WavJEPA [84], which learn representations via self-supervised reconstruction and predictive objectives, respectively. We additionally include AST [85], a vision transformer adapted to audio via supervised training on AudioSet; and CLAP [86], trained via contrastive learning on audio-text pairs. Both of these models have been applied to audio classification tasks beyond speech [62, 87, 88].

Domain-specific Models. Domain-specific models are pretrained on data from a targeted application domain rather than general speech or audio corpora, offering a direct test of whether specialized pretraining confers an advantage on clinical tasks. We include emotion2vec+ [89], pretrained on emotional speech via self-supervised teacher-student distillation; and OPERA [90], a model pretrained on respiratory audio via a multi-task self-supervised objective.

4.3 Evaluation Protocol

To evaluate the performance of pretrained representations across clinical speech tasks, we adopt a linear probing protocol following standard practice [90, 62, 64]. Encoder weights are kept frozen throughout, and a single linear layer is trained on top of the mean-pooled encoder output. The output dimensionality is set to the number of target classes for classification tasks or a single unit for regression. Compared to full fine-tuning, this approach is computationally efficient and less prone to overfitting, making it appropriate for the limited dataset sizes typical in clinical tasks.

Official data splits are used where available; otherwise, datasets are partitioned in one of three ways depending on their size. Datasets with a large number of speakers are split into training (70%), validation (10%), and test (20%) sets with speaker-disjoint partitions stratified by label, sex, and age when possible. Datasets with a small number of speakers are split via 5-fold cross-validation with speaker-disjoint folds stratified by label. Finally, the official split for COVID-19 Sounds [74] is replaced with a custom speaker-disjoint partition due to speaker leakage in the original release.

Beyond standard within-dataset evaluation, we also examine zero-shot cross-condition transfer, in which a linear probe trained on a source dataset is evaluated directly on a disjoint target dataset without any exposure to target data during training. The source dataset is split into training (80%) and validation (20%) sets, and the entire target dataset serves as the test set. Source and target datasets

may belong to the same or different stages in Berisha and Liss' framework. We restrict zero-shot evaluation to the binary classification tasks from the main benchmark. For consistency across datasets, controls are grouped as one class and clinically affected participants as the other; for emotion datasets, non-negative emotions are mapped to controls and negative emotions to the affected class as the closest analogue to the clinical dichotomy.

We evaluate classification tasks using the area under the receiver operating characteristic curve (AUC), reported as the macro-average across classes for multi-class settings. Regression tasks are evaluated using mean absolute error (MAE). For datasets evaluated on a held-out test set, 95% confidence intervals for both AUC and MAE are estimated via bootstrap resampling ($n = 1{,}000$). For datasets evaluated under cross-validation, mean performance across folds is reported. To compare results across tasks with mixed metrics, we also report mean reciprocal rank (MRR) by ranking models on each task within a speech-production stage and averaging the reciprocal of those ranks across all tasks in the stage. Additional implementation details are provided in Appendix D.

4.4 Implementation Details

Embedding extraction for all 12 encoders across the benchmark was performed on compute nodes with 8× NVIDIA H100 80GB GPUs, with extraction parallelized across multiple GPUs and taking approximately 288 GPU-hours total. GPU utilization averaged 70% due to variable-length audio samples, with outlier samples requiring dedicated GPU memory; batching strategies could further optimize this process. Once embeddings were cached, all linear probing experiments (main benchmark, zero-shot transfer, and data efficiency) were executed locally with 8 concurrent jobs, requiring approximately 20 hours of wall-clock time.

5 Results

5.1 Benchmark Evaluation

Figure 1 shows the performance of 12 state-of-the-art audio encoders across all 27 tasks in SpeechDx according to per-task AUC and MAE results with 95% confidence intervals, alongside category-level and overall MRR scores. The results show that the speech-production taxonomy is a meaningful predictor of task difficulty. Conceptualization tasks are consistently difficult, with depression detection (T1) yielding an AUC between 0.40 and 0.65 and depression severity estimation (T2) showing similarly low performance across models. In contrast, tasks belonging to the formulation and neuromuscular articulation stages are more tractable. Aphasia detection (T9) reaches a maximum AUC of 0.97, and the best models exceed an AUC of 0.82 on all neuromuscular articulation classification tasks (T10, T12, T13, T17, T18).

However, performance is not determined by the speech-production stage alone. Although respiratory tasks directly affect breath support and vocal function, performance remains modest across models, with the best AUC reaching 0.79 for COVID-19 detection (T25). In contrast, vocal pathology detection (T27) is substantially more separable, with WavJEPA achieving an AUC of 0.93. This contrast suggests that heterogeneous recording conditions and label noise can attenuate the expected relationship between production-stage proximity and predictive performance. Unlike other datasets, several in the respiratory category are crowdsourced across diverse devices and environments, introducing acquisition variability and confounding factors that obscure the clinical signal.

Overall, no single encoder achieves uniformly strong performance across all conditions impacting speech. The strongest overall models are Whisper (MRR: 0.44), Qwen3-TTS-Tokenizer (MRR: 0.40), and WavLM (MRR: 0.38), whereas CLAP and wav2vec 2.0 are among the weakest performers. The strong aggregate performance of Whisper and Qwen3 suggests an association between scale and broader clinical utility, reflecting their status as the models with the most extensive pretraining. However, overall MRR masks stage-specific variation; emotion2vec+ dominates conceptualization tasks (MRR: 0.77), AST and Whisper perform best on neuromuscular articulation tasks (MRR: 0.60 and 0.44), and Qwen3 and Whisper lead phonatory / respiratory tasks (MRR: 0.54 and 0.53). These discrepancies indicate that current models encode different dimensions of clinical speech variation, but none provides a representation that generalizes reliably across clinical domains.

Domain-specific pretraining yields selective gains. The strong performance of emotion2vec+ on emotion tasks (T3–T6) is likely due to the inclusion of the IEMOCAP dataset in its pretraining

Task	Metric	General Speech						General Audio				Domain-specific	
		w2v2	HuBERT	WavLM	MMS	Qwen3	Whisper	AudioMAE	WavJEPA	AST	CLAP	emo2vec+	OPERA
Conceptualization													
T1 (Depression detection)	AUC	0.40 (0.22, 0.58)	0.59 (0.41, 0.75)	0.57 (0.40, 0.73)	0.55 (0.38, 0.70)	0.46 (0.29, 0.63)	0.55 (0.37, 0.72)	0.57 (0.41, 0.72)	0.57 (0.40, 0.74)	**0.65** (0.47, 0.82)	0.60 (0.45, 0.74)	0.64 (0.46, 0.80)	0.43 (0.28, 0.60)
T2 (Depression severity)	MAE ↓	5.46 (4.48, 6.50)	5.45 (4.46, 6.50)	5.37 (4.40, 6.37)	5.46 (4.46, 6.53)	5.31 (4.30, 6.38)	5.34 (4.35, 6.35)	**5.27** (4.31, 6.30)	5.43 (4.44, 6.50)	5.35 (4.36, 6.39)	5.41 (4.44, 6.45)	5.40 (4.44, 6.43)	5.33 (4.34, 6.38)
T3 (Emotion classification)	AUC	0.56 (0.50, 0.63)	0.79 (0.76, 0.83)	0.86 (0.84, 0.88)	0.77 (0.75, 0.78)	0.89 (0.86, 0.91)	0.87 (0.85, 0.90)	0.82 (0.79, 0.85)	0.83 (0.81, 0.85)	0.87 (0.86, 0.89)	0.65 (0.64, 0.66)	**0.98** (0.97, 0.99)	0.71 (0.67, 0.74)
T4 (Binary emotion classification)	AUC	0.60 (0.58, 0.63)	0.73 (0.69, 0.76)	0.76 (0.75, 0.78)	0.63 (0.57, 0.70)	0.78 (0.75, 0.81)	0.82 (0.79, 0.84)	0.71 (0.67, 0.74)	0.73 (0.68, 0.78)	0.77 (0.74, 0.81)	0.60 (0.57, 0.63)	**0.98** (0.96, 0.99)	0.62 (0.58, 0.65)
T5 (Emotion classification)	AUC	0.68 (0.65, 0.70)	0.78 (0.74, 0.82)	0.83 (0.81, 0.85)	0.79 (0.76, 0.82)	0.83 (0.81, 0.86)	0.86 (0.85, 0.88)	0.80 (0.76, 0.83)	0.78 (0.77, 0.80)	0.80 (0.79, 0.82)	0.68 (0.66, 0.70)	**0.92** (0.90, 0.93)	0.74 (0.73, 0.75)
T6 (Binary emotion classification)	AUC	0.64 (0.60, 0.67)	0.69 (0.66, 0.72)	0.76 (0.73, 0.80)	0.64 (0.62, 0.67)	0.78 (0.75, 0.81)	0.76 (0.73, 0.79)	0.70 (0.67, 0.73)	0.68 (0.65, 0.71)	0.71 (0.68, 0.73)	0.58 (0.56, 0.59)	**0.89** (0.88, 0.89)	0.62 (0.60, 0.64)
Mean Reciprocal Rank	MRR	0.09	0.15	0.22	0.11	0.38	0.31	0.29	0.15	0.36	0.13	**0.77**	0.14
Formulation													
T7 (Alzheimer's detection)	AUC	0.65 (0.49, 0.80)	0.58 (0.40, 0.74)	0.69 (0.53, 0.83)	0.62 (0.45, 0.77)	0.74 (0.58, 0.88)	**0.75** (0.59, 0.89)	0.59 (0.42, 0.75)	0.66 (0.47, 0.81)	0.47 (0.29, 0.66)	0.36 (0.20, 0.53)	0.64 (0.47, 0.80)	0.53 (0.35, 0.71)
T8 (Alzheimer's severity)	MAE ↓	6.86 (5.92, 7.82)	6.94 (5.95, 7.90)	8.58 (7.23, 9.87)	6.75 (5.79, 7.70)	10.75 (9.25, 12.20)	9.97 (8.60, 11.33)	6.72 (5.84, 7.73)	9.52 (8.26, 10.81)	8.14 (6.93, 9.36)	6.96 (5.98, 7.96)	6.83 (5.86, 7.84)	**6.09** (5.27, 6.94)
T9 (Aphasia detection)	AUC	0.92 (0.85, 0.97)	0.92 (0.86, 0.97)	0.96 (0.91, 0.99)	0.93 (0.86, 0.98)	**0.97** (0.92, 1.00)	0.92 (0.84, 0.98)	0.90 (0.80, 0.97)	0.92 (0.85, 0.98)	0.95 (0.89, 0.99)	0.83 (0.68, 0.94)	0.67 (0.53, 0.79)	0.94 (0.87, 0.99)
Mean Reciprocal Rank	MRR	0.17	0.15	0.31	0.23	**0.53**	0.41	0.24	0.16	0.18	0.11	0.17	0.45
Articulation (Neuromuscular)													
T10 (Dysarthria detection)	AUC	0.75 (0.62, 0.88)	0.88 (0.77, 0.99)	0.88 (0.76, 1.00)	0.75 (0.57, 0.93)	0.87 (0.77, 0.97)	**0.91** (0.84, 0.98)	0.83 (0.67, 0.98)	0.73 (0.53, 0.93)	0.70 (0.46, 0.93)	0.75 (0.60, 0.89)	0.79 (0.71, 0.86)	0.66 (0.44, 0.88)
T11 (Dysarthria severity)	MAE ↓	0.67 (0.49, 0.85)	0.51 (0.35, 0.67)	0.63 (0.37, 0.90)	0.53 (0.20, 0.87)	0.54 (0.33, 0.75)	**0.43** (0.27, 0.59)	0.56 (0.37, 0.75)	0.57 (0.23, 0.92)	0.48 (0.29, 0.68)	0.60 (0.41, 0.80)	0.69 (0.56, 0.81)	0.79 (0.45, 1.14)
T12 (Dysarthria detection)	AUC	0.83 (0.72, 0.94)	0.94 (0.87, 1.00)	0.96 (0.91, 1.00)	**0.99** (0.98, 0.99)	0.95 (0.90, 0.99)	0.97 (0.92, 1.00)	0.94 (0.88, 1.00)	0.95 (0.90, 1.00)	0.97 (0.94, 1.00)	0.90 (0.84, 0.96)	0.82 (0.69, 0.94)	0.94 (0.91, 0.98)
T13 (Parkinson's detection)	AUC	0.52 (0.37, 0.67)	0.60 (0.53, 0.66)	0.81 (0.75, 0.88)	0.73 (0.59, 0.86)	0.67 (0.51, 0.83)	0.73 (0.65, 0.81)	0.81 (0.64, 0.98)	0.66 (0.49, 0.84)	**0.82** (0.74, 0.91)	0.73 (0.53, 0.93)	0.74 (0.58, 0.89)	0.80 (0.75, 0.85)
T14 (Parkinson's severity)	MAE ↓	0.46 (0.43, 0.50)	0.46 (0.43, 0.49)	0.44 (0.42, 0.47)	0.46 (0.43, 0.49)	0.42 (0.36, 0.49)	0.43 (0.39, 0.46)	0.44 (0.39, 0.50)	0.47 (0.44, 0.50)	**0.33** (0.26, 0.41)	0.47 (0.43, 0.51)	0.41 (0.36, 0.46)	0.38 (0.33, 0.43)
T15 (Parkinson's severity)	MAE ↓	0.46 (0.38, 0.53)	0.46 (0.41, 0.51)	0.44 (0.38, 0.49)	0.46 (0.39, 0.52)	0.42 (0.33, 0.51)	0.44 (0.39, 0.49)	0.43 (0.37, 0.50)	0.46 (0.39, 0.53)	**0.35** (0.27, 0.42)	0.46 (0.41, 0.51)	0.42 (0.36, 0.49)	0.39 (0.35, 0.43)
T16 (Parkinson's severity)	MAE ↓	0.47 (0.45, 0.50)	0.47 (0.44, 0.49)	0.45 (0.41, 0.48)	0.46 (0.42, 0.50)	0.41 (0.34, 0.48)	0.44 (0.41, 0.46)	0.43 (0.38, 0.48)	0.48 (0.45, 0.52)	**0.34** (0.30, 0.38)	0.48 (0.44, 0.51)	0.41 (0.37, 0.45)	0.39 (0.33, 0.45)
T17 (Disfluency detection)	AUC	0.78 (0.73, 0.83)	0.84 (0.81, 0.88)	**0.86** (0.83, 0.88)	0.73 (0.63, 0.83)	0.81 (0.75, 0.86)	0.85 (0.81, 0.89)	0.72 (0.63, 0.81)	0.76 (0.73, 0.79)	0.74 (0.65, 0.83)	0.53 (0.37, 0.68)	0.59 (0.55, 0.64)	0.59 (0.49, 0.70)
T18 (Disfluency classification)	AUC	0.71 (0.69, 0.74)	0.79 (0.77, 0.80)	**0.84** (0.82, 0.85)	0.76 (0.74, 0.78)	0.78 (0.78, 0.79)	0.81 (0.78, 0.83)	0.74 (0.72, 0.76)	0.75 (0.72, 0.77)	0.73 (0.70, 0.76)	0.54 (0.52, 0.56)	0.59 (0.58, 0.61)	0.61 (0.59, 0.63)
Mean Reciprocal Rank	MRR	0.12	0.21	0.43	0.25	0.23	0.44	0.18	0.13	**0.60**	0.11	0.18	0.25
Articulation (Phonatory/Respiratory)													
T19 (Respiratory symptom detection)	AUC	0.56 (0.53, 0.59)	0.63 (0.60, 0.67)	0.68 (0.65, 0.71)	0.65 (0.63, 0.68)	**0.69** (0.66, 0.72)	0.68 (0.65, 0.71)	0.61 (0.58, 0.64)	0.63 (0.60, 0.66)	0.63 (0.60, 0.66)	0.59 (0.56, 0.62)	0.57 (0.54, 0.61)	0.62 (0.58, 0.65)
T20 (Respiratory symptom detection)	AUC	0.56 (0.55, 0.57)	0.60 (0.58, 0.61)	0.65 (0.64, 0.66)	0.62 (0.61, 0.63)	0.64 (0.62, 0.65)	**0.65** (0.64, 0.66)	0.59 (0.58, 0.61)	0.61 (0.59, 0.62)	0.60 (0.58, 0.61)	0.55 (0.53, 0.56)	0.57 (0.55, 0.58)	0.57 (0.55, 0.58)
T21 (COVID-19 detection)	AUC	0.60 (0.51, 0.68)	0.61 (0.52, 0.69)	**0.65** (0.56, 0.72)	0.54 (0.45, 0.64)	0.61 (0.52, 0.69)	0.61 (0.53, 0.69)	0.47 (0.39, 0.54)	0.53 (0.43, 0.61)	0.48 (0.41, 0.56)	0.55 (0.47, 0.64)	0.60 (0.52, 0.68)	0.50 (0.41, 0.59)
T22 (COVID-19 detection)	AUC	0.64 (0.59, 0.69)	0.66 (0.62, 0.71)	0.68 (0.63, 0.73)	0.69 (0.64, 0.74)	0.65 (0.59, 0.70)	**0.70** (0.65, 0.74)	0.62 (0.57, 0.67)	0.66 (0.61, 0.71)	0.64 (0.60, 0.68)	0.58 (0.54, 0.62)	0.54 (0.49, 0.59)	0.64 (0.59, 0.68)
T23 (Respiratory symptom classification)	AUC	0.55 (0.54, 0.56)	0.60 (0.59, 0.60)	0.60 (0.59, 0.61)	0.60 (0.59, 0.61)	0.60 (0.59, 0.61)	**0.61** (0.61, 0.62)	0.57 (0.56, 0.58)	0.59 (0.58, 0.60)	0.58 (0.57, 0.59)	0.55 (0.54, 0.56)	0.56 (0.55, 0.57)	0.57 (0.56, 0.57)
T24 (Respiratory symptom detection)	AUC	0.65 (0.61, 0.69)	0.71 (0.67, 0.75)	0.72 (0.68, 0.76)	0.66 (0.62, 0.70)	**0.73** (0.69, 0.77)	0.72 (0.68, 0.76)	0.68 (0.64, 0.73)	0.66 (0.61, 0.70)	0.69 (0.64, 0.73)	0.60 (0.55, 0.65)	0.66 (0.62, 0.71)	0.64 (0.59, 0.69)
T25 (COVID-19 detection)	AUC	0.74 (0.69, 0.79)	0.74 (0.69, 0.79)	0.77 (0.72, 0.81)	0.73 (0.68, 0.77)	**0.79** (0.74, 0.83)	0.77 (0.72, 0.81)	0.74 (0.70, 0.79)	0.70 (0.65, 0.75)	0.75 (0.70, 0.80)	0.61 (0.55, 0.66)	0.69 (0.65, 0.74)	0.67 (0.62, 0.71)
T26 (Respiratory symptom classification)	AUC	0.56 (0.53, 0.60)	0.58 (0.55, 0.62)	0.59 (0.56, 0.63)	0.57 (0.54, 0.60)	0.60 (0.57, 0.63)	0.59 (0.55, 0.62)	0.58 (0.54, 0.61)	**0.60** (0.57, 0.64)	0.60 (0.57, 0.65)	0.55 (0.51, 0.58)	0.56 (0.53, 0.60)	0.60 (0.56, 0.63)
T27 (Vocal pathology detection)	AUC	0.65 (0.60, 0.71)	0.72 (0.67, 0.78)	0.91 (0.87, 0.95)	0.91 (0.87, 0.95)	0.92 (0.87, 0.95)	0.81 (0.75, 0.86)	0.89 (0.84, 0.93)	**0.93** (0.89, 0.97)	0.91 (0.86, 0.95)	0.70 (0.61, 0.77)	0.67 (0.61, 0.72)	0.89 (0.83, 0.94)
Mean Reciprocal Rank	MRR	0.11	0.20	0.47	0.25	**0.54**	0.53	0.13	0.34	0.20	0.09	0.11	0.13
Overall													
Mean Reciprocal Rank	MRR	0.11	0.18	0.38	0.21	0.40	**0.44**	0.20	0.21	0.37	0.11	0.29	0.21

Figure 1: The benchmark evaluation of 12 audio encoders across 27 clinical speech AI tasks, grouped according to Berisha and Liss' framework of speech production [26]. Classification tasks report AUC and regression tasks report MAE, with 95% bootstrap confidence intervals for tasks evaluated with a held-out test set or (μ - σ, μ + σ) for tasks evaluated with 5-fold cross-validation. MRR is computed within each category and overall. Model names are shortened as follows: w2v2 = wav2vec 2.0, Qwen3 = Qwen3-TTS-Tokenizer, emo2vec+ = emotion2vec+, and OPERA = OPERA-GT.

corpus [89]. Its poor performance on depression detection, a closely related conceptualization-stage condition, illustrates that domain-specific pretraining does not generalize to even clinically proximate tasks. OPERA, another domain-specific model, shows a similar mismatch. Although pretrained

on respiratory audio, it performs weakly on respiratory tasks (MRR: 0.13) and is instead most competitive on formulation tasks (MRR: 0.45). These patterns suggest that narrow domain alignment between pretraining and downstream tasks is insufficient for general clinical speech understanding. Finally, to assess performance under data scarcity, a central constraint in clinical settings, we conduct data efficiency experiments varying training set size from 12.5% to 100% for the top three models, with results reported in Appendix E.

5.2 Zero-shot Cross-condition Transfer

To assess whether the representations learned by current audio encoders capture clinical structure beyond dataset-specific artefacts, we evaluate zero-shot transfer between datasets. Figure 2 reports the best zero-shot AUC and the corresponding model for each source–target pair; for cross-category transfer, source and target datasets are pooled within each stage. Complete results for all models and all transfer pairs are provided in Appendix F.

Within-category transfer. For many source–target pairs within a production stage, the best zero-shot transfer approaches within-dataset performance, with several cases meeting or exceeding it. A probe trained on DementiaBank reaches an AUC of 0.94 on aphasia detection (HuBERT) compared to a within-dataset best of 0.97, and probes trained on either emotion dataset reach AUCs of 0.74–0.75 on depression detection (emotion2vec+) compared to a within-dataset best of 0.65. Transfer is consistently weaker across phonatory / respiratory tasks (AUC: 0.57–0.69), identifying this category as the most resistant to cross-dataset generalization. Notably, the best zero-shot model differs from the best within-dataset model in approximately half of all source–target pairs. For example, AudioMAE exceeds the best within-dataset performance on Parkinson's detection when trained on UASpeech, and Whisper leads dysarthria-to-dysarthria transfer over the within-dataset leader MMS, indicating that within-dataset and zero-shot regimes reward different representational properties.

Cross-condition transfer. Across production stages, transfer is variable but reveals a clear asymmetry. Representations trained on phonatory / respiratory data transfer into conceptualization and formulation tasks at AUCs of 0.83 (emotion2vec+) and 0.88 (HuBERT) respectively, while transfer in the reverse direction does not exceed an AUC of 0.60. Formulation also transfers effectively into neuromuscular tasks (AUC: 0.80, AST). Notably, model rankings observed in the within-dataset evaluation do not predict cross-category transfer performance. Qwen3 achieves the highest within-dataset performance on phonatory / respiratory tasks but fails to transfer across categories, while emotion2vec+, which shows mid-tier within-dataset performance leads cross-category transfer to phonatory / respiratory targets. Progress on clinical speech AI will require representations that maintain performance across regimes, rather than models that excel within a single dataset; closing this gap remains an open challenge for future work.

6 Broader Impact and Limitations

Speech-based health assessment could transform everyday devices into diagnostic tools, extending healthcare access to underserved populations. The ubiquity of microphones available in smartphones, smart speakers, and even earbuds enables frequent monitoring with minimal user effort. These opportunities not only circumvent the logistical barriers inherent in traditional clinical assessments but also enable the capture of subtle physiological shifts.

Realizing the full diagnostic potential of clinical speech AI requires representations that generalize across conditions, populations, and recording settings. In pursuit of this objective, SpeechDx provides a unified evaluation infrastructure to drive progress toward clinically deployable models while ensuring that performance gains are meaningful rather than driven by dataset idiosyncrasies.

One key consideration for SpeechDx is the utility and objectivity of the underlying labels for each dataset. The majority of tasks are designed to reveal acoustic markers that classify people with and without a medical condition. The resulting markers would ideally be used to screen people with nascent symptoms; however, these datasets typically comprise people who are either completely healthy or have lived with the condition for multiple years. These datasets also represent single assessments rather than disease progression, though many conditions vary significantly over time. Several other tasks rely on patients' or clinicians' answers to standardized questionnaires (e.g.,

Figure 2: The evaluation of zero-shot transfer for classification tasks. The top four grids show transfer across tasks within the same category, while the bottom grid shows averaged cross-category performance. Each cell reports the best model and the AUC it achieved for the given task.

PHQ-8 [91], UPDRS [92]). Although the questionnaires are clinically validated and human-reported responses should not be discredited, there is unavoidable subjectivity associated with anchoring bias, recall bias, and person-specific internal scaling that may obscure physiological information [93, 94].

Another consideration relates to the diversity of human subjects across the benchmark. SpeechDx primarily comprises English-language recordings, though diverse regional accents are represented through datasets collected outside North America. We did not conduct a formal analysis of how this factor might have confounded our zero-shot analysis, so this remains an opportunity for future work. Sex and age are other demographic factors that warrant further investigation, particularly since certain populations are predisposed to specific conditions (e.g., elderly women and dementia [95]).

Lastly, SpeechDx does not exhaustively cover the clinical domains in which speech-based assessment has shown promise. Conditions such as Huntington's disease [96], pediatric speech sound disorders [97], and chronic obstructive pulmonary disease [98, 36] are a subset of the diverse clinical domains where vocal biomarkers have shown promise, yet publicly available datasets are either unavailable or not sizable enough for thorough evaluation. Subsequent iterations will incorporate these and other emerging areas to provide an increasingly comprehensive evaluation of clinical speech AI.

7 Conclusion

SpeechDx establishes a comprehensive evaluation framework for clinical speech AI. While large-scale general-purpose speech models currently provide the most robust baselines, experiments with our benchmark reveal that no single representation yet generalizes reliably across the diverse clinical landscape. Moreover, cross-condition transfer analyses demonstrate critical gaps in model robustness and data efficiency. Ultimately, SpeechDx serves as a foundation for tracking progress toward truly generalizable representations capable of supporting real-world clinical monitoring.

References

[1] Vesna Despotovic, Mohamad Ismael, Marc Cornil, Romain M. Call, and Guy Fagherazzi. Detection of covid-19 from voice, cough and breathing patterns: Dataset and preliminary results. *Computer Biology and Medicine*, 138:104944, 2021.

[2] Mohammed Usman, Vinit Kumar Gunjan, Mohd Wajid, Mohammed Zubair, and Kazy Noor-e-alam Siddiquee. Speech as a biomarker for covid-19 detection using machine learning. *Computational Intelligence and Neuroscience*, 2022(1):6093613, 2022.

[3] David H. Shih, Chih-Hao Liao, Tzu-Wei Wu, Xiao-Yu Xu, and Ming-Hsiang Shih. Dysarthria speech detection using convolutional neural networks with gated recurrent unit. *Healthcare*, 10(10):1956, 2022.

[4] Carlos D. Ríos-Urrego, Jan Rusz, Elmar Nöth, and Juan R. Orozco-Arroyave. Automatic classification of hypokinetic and hyperkinetic dysarthria based on gmm-supervectors. In *Proceedings of INTERSPEECH 2023*. ISCA, 2023.

[5] Farhad Javanmardi, Sudarsana Reddy Kadiri, and Paavo Alku. Pre-trained models for detection and severity level classification of dysarthria from speech. *Speech Communication*, 158:103047, 2024.

[6] Mittapalle Kiran Reddy and Paavo Alku. Exemplar-based sparse representations for detection of parkinson's disease from speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1386–1396, 2023.

[7] Rania Khaskhoussy and Yosra Ben Ayed. Improving parkinson's disease recognition through voice analysis using deep learning. *Pattern Recognition Letters*, 168:64–70, 2023.

[8] Laura Moro-Velazquez, Juan A. Gomez-Garcia, Juan D. Arias-Londoño, Najim Dehak, and Juan I. Godino-Llorente. Advances in parkinson's disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects. *Biomedical Signal Processing and Control*, 66:102418, 2021.

[9] Mahsa Zolnoori, Arash Zolnour, and Maxim Topaz. Adscreen: A speech processing-based screening system for automatic identification of patients with alzheimer's disease and related dementia. *Artificial Intelligence in Medicine*, 143:102624, 2023.

[10] Israel Martínez-Nicolás, Thide E Llorente, Francisco Martínez-Sánchez, and Juan José G Meilán. Ten years of research on automatic voice and speech analysis of people with alzheimer's disease and mild cognitive impairment: a systematic review article. *Frontiers in Psychology*, 12:620251, 2021.

[11] Felix Braun, Maria Schuster, Florian Honig, Elmar Noeth, and Juan Rafael Orozco-Arroyave. Classifying dementia in the presence of depression: A cross-corpus study. In *Proceedings of INTERSPEECH 2023*. ISCA, 2023.

[12] Pingping Wu, Ruihao Wang, Han Lin, Fanlong Zhang, Juan Tu, and Miao Sun. Automatic depression recognition by intelligent speech signal processing: A systematic survey. *CAAI Transactions on Intelligence Technology*, 8(3):701–711, 2023.

[13] Sanne Koops, Sanne G Brederoo, Janna N De Boer, Femke G Nadema, Alban E Voppel, and Iris E Sommer. Speech as a biomarker for depression. *CNS & Neurological Disorders-Drug Targets-CNS & Neurological Disorders)*, 22(2):152–160, 2023.

[14] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. A review of depression and suicide risk assessment using speech analysis. *Speech communication*, 71:10–49, 2015.

[15] Guo-Shiang Liu, Nikola Jovanovic, Chang K. Sung, and Philip C. Doyle. A scoping review of artificial intelligence detection of voice pathology: Challenges and opportunities. *Otolaryngology–Head and Neck Surgery*, 171(3):658–666, 2024.

[16] Alkis Koudounas, Moreno La Quatra, Gabriele Ciravegna, Marco Fantini, Erika Crosetti, Giovanni Succo, Tania Cerquitelli, Sabato Marco Siniscalchi, and Elena Baralis. MVP: Multi-source Voice Pathology detection. In *Interspeech 2025*, pages 3548–3552, 2025.

[17] Björn Schuller, Anton Batliner, Shahin Amiriparian, Christian Bergler, Maurice Gerczuk, Natalie Holz, Pauline Larrouy-Maestri, Sebastien Bayerl, Korbinian Riedhammer, Adria Mallol-Ragolta, Maria Pateraki, Harry Coppock, Ivan Kiskin, Marianne Sinka, and Stephen Roberts. The acm multimedia 2022 computational paralinguistics challenge: Vocalisations, stuttering, activity, & mosquitoes. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 7120–7124, New York, NY, USA, 2022. Association for Computing Machinery.

[18] Richard C. Gale, Megan Fleegle, Gerasimos Fergadiotis, and Steven Bedrick. The post-stroke speech transcription (psst) challenge. In *Proceedings of the LREC 2022 RaPID-4 Workshop*, pages 41–55, 2022.

[19] Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge. In *Interspeech 2020*, pages 2172–2176, 2020.

[20] Wen Wu, Ziyun Cui, Chang Lei, Yinan Duan, Diyang Qu, Ji Wu, Bowen Zhou, Runsen Chen, and Chao Zhang. The 1st speechwellness challenge: Detecting suicide risk among adolescents. In *Interspeech 2025*, pages 399–403. ISCA, 2025.

[21] Ananya Muguli, Lancelot Pinto, Nirmala R., Neeraj Sharma, Prashant Krishnan, Prasanta Kumar Ghosh, Rohit Kumar, Shrirama Bhat, Srikanth Raj Chetupalli, Sriram Ganapathy, Shreyas Ramoji, and Viral Nanda. Dicova challenge: Dataset, task, and baseline system for covid-19 diagnosis using acoustics, 2021.

[22] Saturnino Luz, Fasih Haider, Davida Fromm, Ioulietta Lazarou, Ioannis Kompatsiaris, and Brian MacWhinney. Multilingual alzheimer's dementia recognition through spontaneous speech: a signal processing grand challenge, 2023.

[23] Visar Berisha, Chelsea Krantsevich, Gabriela Stegmann, Shira Hahn, and Julie Liss. Are reported accuracies in the clinical speech machine learning literature overoptimistic? In *Proceedings of INTERSPEECH 2022*, pages 2453–2457. ISCA, 09 2022.

[24] Guilherme Schu, Parvaneh Janbakhshi, and Ina Kodrasi. On using the ua-speech and torgo databases to validate automatic dysarthric speech classification approaches. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2022.

[25] Yi Zhu, Mohamed Imoussaine, Carolyn Côté-Lussier, and Tiago Falk. Investigating biases in covid-19 diagnostic systems processed with automated speech anonymization algorithms. pages 46–54, 08 2023.

[26] Visar Berisha and Julie Liss. Responsible development of clinical speech ai: Bridging the gap between clinical research and technology. *npj Digital Medicine*, 7, 08 2024.

[27] Katherine Verdolini, Clark A. Rosen, and Ryan C. Branski, editors. *Classification Manual for Voice Disorders-I*. Psychology Press, 1 edition, 2006.

[28] Guy Fagherazzi, Aurélie Fischer, Muhannad Ismael, and Vladimir Despotovic. Voice for health: the use of vocal biomarkers from research to clinical practice. *Digital biomarkers*, 5(1):78–88, 2021.

[29] Jessica Robin, John E. Harrison, Liam D. Kaufman, Frank Rudzicz, William Simpson, and Maria Yancheva. Evaluation of speech-based digital biomarkers: Review and recommendations. *Digital Biomarkers*, 4(3):99–108, 10 2020.

[30] Vikram Ramanarayanan, Adam Lammert, Hannah Rowe, Thomas Quatieri, and Jordan Green. Speech as a biomarker: Opportunities, interpretability, and challenges. *Perspectives of the ASHA Special Interest Groups*, 7:276–283, 01 2022.

[31] Si-Ioi Ng, Lingfeng Xu, Ingo Siegert, Nicholas Cummins, Nina R Benway, Julie Liss, and Visar Berisha. An end-to-end overview of clinical speech ai. *IEEE Transactions on Audio, Speech and Language Processing*, 34:1016–1048, 2026.

[32] Florian Eyben, Martin Wollmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462, New York, NY, USA, 2010. Association for Computing Machinery.

[33] Yi Zhu, Abhishek Tiwari, João Monteiro, Shruti Kshirsagar, and Tiago Henrique Falk. Covid-19 detection via fusion of modulation spectrum and linear prediction speech features. *IEEE/ACM transactions on audio, speech, and language processing*, 31:1536–1549, 2023.

[34] Tomas Arias-Vergara, Juan Camilo Vasquez-Correa, and Juan Rafael Orozco-Arroyave. Parkinson's disease and aging: Analysis of their effect in phonation and articulation of speech. *Cognitive Computation*, 9(6):731–748, 2017.

[35] Ahmed Afshan, Jian Guo, Seong Joon Park, Venkatesh Ravi, Jonathan Flint, and Abeer Alwan. Effectiveness of voice quality features in detecting depression. In *Proceedings of INTERSPEECH 2018*, pages 1676–1680. ISCA, 2018.

[36] Sejal Bhalla, Deshang Kong, Salaar Liaqat, Daniyal Liaqat, Robert Wu, Andrea Gershon, Eyal de Lara, and Alex Mariakakis. Association of daily lung condition in copd patients with wearable speech and physiological data. *Scientific reports.*, 15(1), 2025-12-29.

[37] Juan Camilo Vasquez-Correa, Juan Rafael Orozco-Arroyave, Tobias Bocklet, and Elmar Noeth. Towards an automatic evaluation of the dysarthria level of patients with parkinson's disease. *Journal of Communication Disorders*, 76:21–36, 2018.

[38] Chunying Fang, Haifeng Li, Lin Ma, and Mancai Zhang. Intelligibility evaluation of pathological speech through multigranularity feature extraction and optimization. *Computational and Mathematical Methods in Medicine*, 2017(1):2431573, 2017.

[39] Nicholas Cummins, Alice Baird, and Björn W. Schuller. Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods*, 151:41–54, 2018. Health Informatics and Translational Data Analytics.

[40] Qin Yang, Xin Li, Xinyun Ding, Feiyang Xu, and Zhenhua Ling. Deep learning-based speech analysis for alzheimer's disease detection: A literature review. *Alzheimer's Research & Therapy*, 14(1):186, 2022.

[41] Rumana Islam, Esam Abdel-Raheem, and Mohammed Tarique. Voice pathology detection using convolutional neural networks with electroglottographic (egg) and speech signals. *Computer Methods and Programs in Biomedicine Update*, 2:100074, 2022.

[42] Harry Coppock, Alex Gaskell, Panagiotis Tzirakis, Alice Baird, Lyn Jones, and Björn Schuller. End-to-end convolutional neural network enables covid-19 detection from breath and cough audio: a pilot study. *BMJ Innovations*, 7(2):356–362, 2021.

[43] Herath Mudiyanselage Dhammike Piyumal Madhurajith Herath, Weraniyagoda Arachchilage Sahanaka Anuththara Weraniyagoda, Rajapakshage Thilina Madhushan Rajapaksha, Patikiri Arachchige Don Shehan Nilmantha Wijesekara, Kalupahana Liyanage Kushan Sudheera, and Peter Han Joo Chong. Automatic assessment of aphasic speech sensed by audio sensors for classification into aphasia severity levels to recommend speech therapies. *Sensors*, 22(18), 2022.

[44] Ah Young Kim, Eun Hye Jang, Seung-Hwan Lee, Kwang-Yeon Choi, Jeon Gue Park, and Hyun-Chool Shin. Automatic depression detection using smartphone-based text-dependent speech signals: Deep convolutional neural network approach. *J Med Internet Res*, 25:e34474, Jan 2023.

[45] Yi Zhu, Alex Mariakakis, Eyal De Lara, and Tiago H. Falk. How generalizable and interpretable are speech-based covid-19 detection systems?: A comparative analysis and new system proposal. In *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–5, 2022.

[46] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

[47] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

[48] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture, 2023.

[49] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[50] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

[51] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[52] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

[53] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021.

[54] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, October 2022.

[55] Farhad Javanmardi, Sudarsana Reddy Kadiri, and Paavo Alku. Pre-trained models for detection and severity level classification of dysarthria from speech. *Speech Communication*, 158:103047, 2024.

[56] Giulia Sanguedolce, Sophie Brook, Dragos C. Gruia, Patrick A. Naylor, and Fatemeh Geranmayeh. When Whisper Listens to Aphasia: Advancing Robust Post-Stroke Speech Recognition. In *Interspeech 2024*, pages 1995–1999, 2024.

[57] Bubai Maji, Shazia Nasreen, Rajlakshmi Guha, Aurobinda Routray, Debabrata Majumdar, and Km Poonam. Exploring self-supervised models for depressive disorder detection: A study on speech corpora. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–4, 2024.

[58] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. Emotion Recognition from Speech Using wav2vec 2.0 Embeddings. In *Interspeech 2021*, pages 3400–3404, 2021.

[59] Zhengjun Yue, Devendra Kayande, Zoran Cvetkovic, and Erfan Loweimi. Probing whisper for dysarthric speech in detection and assessment, 2025.

[60] Xing-Yu Chen, Qiu-Shi Zhu, Jie Zhang, and Li-Rong Dai. Supervised and self-supervised pretraining based covid-19 detection using acoustic breathing/cough/speech signals. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 561–565. IEEE, May 2022.

[61] Yi Zhu and Tiago Falk. Wavrx: A disease-agnostic, generalizable, and privacy-preserving speech health diagnostic model. *IEEE Journal of Biomedical and Health Informatics*, 29(9):6353–6365, 2025.

[62] Sebastien Baur, Zaid Nabulsi, Wei-Hung Weng, Jake Garrison, Louis Blankemeier, Sam Fishman, Christina Chen, Sujay Kakarmath, Minyoi Maimbolwa, Nsala Sanjase, Brian Shuma, Yossi Matias, Greg S. Corrado, Shwetak Patel, Shravya Shetty, Shruthi Prabhakara, Monde Muyoyeta, and Diego Ardila. Hear – health acoustic representations, 2024.

[63] Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. SUPERB: Speech Processing Universal PERformance Benchmark. In *Interspeech 2021*, pages 1194–1198, 2021.

[64] Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W. Schuller, Christian J. Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, Max Henry, Nicolas Pinto, Camille Noufi, Christian Clough, Dorien Herremans, Eduardo Fonseca, Jesse Engel, Justin Salamon, Philippe Esling, Pranay Manocha, Shinji Watanabe, Zeyu Jin, and Yonatan Bisk. Hear: Holistic evaluation of audio representations, 2022.

[65] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. The distress analysis interview corpus of human and computer interviews. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[66] Steven R. Livingstone and Frank A. Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):e0196391, 2018.

[67] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359, 12 2008.

[68] Saturnino Luz, Fasih Haider, Davida Fromm, Ioulietta Lazarou, Ioannis Kompatsiaris, and Brian MacWhinney. Multilingual alzheimer's dementia recognition through spontaneous speech: a signal processing grand challenge. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2. IEEE, 2023.

[69] Margaret M. Forbes, Davida Fromm, and Brian MacWhinney. Aphasiabank: A resource for clinicians. *Aphasiology*, 26(11):1281–1295, 2012.

[70] Frank Rudzicz, Aravind Namasivayam, and Talya Wolff. The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, 46:1–19, 01 2010.

[71] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas S. Huang, Kenneth Watkin, and Simone Frame. Dysarthric speech database for universal access research. In *Interspeech 2008*, pages 1741–1744, 2008.

[72] Hagen Jaeger, Dhaval Trivedi, and Michael Stadtschnitzer. Mobile device voice recordings at king's college london (mdvr-kcl) from both early and advanced parkinson's disease patients and healthy controls. *Zenodo*, 2019.

[73] Sebastian Peter Bayerl, Alexander Wolff von Gudenberg, Florian Hönig, Elmar Noeth, and Korbinian Riedhammer. Ksof: The kassel state of fluency dataset – a therapy centered dataset of stuttering. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1780–1787, Marseille, France, June 2022. European Language Resources Association.

[74] Tong Xia, Dimitris Spathis, Chloe Brown, Jagmohan Chauhan, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Erika Bondareva, Ting Dang, Andres Floto, Pietro Cicuta, and Cecilia Mascolo. Covid-19 sounds: A large-scale audio dataset for digital respiratory screening. In *Proceedings of the 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[75] Debarpan Bhattacharya, Neeraj Kumar Sharma, Debottam Dutta, Srikanth Raj Chetupalli, Pravin Mote, Sriram Ganapathy, C. Chandrakiran, Sahiti Nori, K. K. Suhail, Sadhana Gonuguntla, and Murali Alagesan. Coswara: A respiratory sounds and symptoms dataset for remote screening of sars-cov-2 infection. *Scientific Data*, 10(1):397, 2023.

[76] Luis M.T. Jesus, Inês Belo, Jessica Machado, and Andreia Hall. The advanced voice function assessment databases (avfad): Tools for voice clinicians and speech research. In *Advances in Speech-language Pathology*, chapter 14. IntechOpen, London, 2017.

[77] Lester Phillip Violeta, Wen-Chin Huang, and Tomoki Toda. Investigating self-supervised pretraining frameworks for pathological speech recognition, 2022.

[78] Jie Cai, Yuliang Song, Jianghao Wu, and Xiong Chen. Voice disorder classification using wav2vec 2.0 feature extraction. *Journal of Voice*, 2024.

[79] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling speech technology to 1,000+ languages, 2023.

[80] Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.

[81] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[82] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[83] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen, 2023.

[84] Goksenin Yuksel, Pierre Guetschel, Michael Tangermann, Marcel van Gerven, and Kiki van der Heijden. Wavjepa: Semantic learning unlocks robust audio foundation models for raw waveforms. *arXiv preprint arXiv:2509.23238*, 2025.

[85] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575, 2021.

[86] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation, 2022.

[87] Mohammad Tami, Sari Masri, Ahmad Hasasneh, and Chakib Tadj. Transformer-based approach to pathology diagnosis using audio spectrogram. *Information*, 15(5), 2024.

[88] Nuwan Madusanka and Byeong il Lee. Vocal biomarkers for parkinson's disease classification using audio spectrogram transformers. *Journal of Voice*, 2024.

[89] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*, 2023.

[90] Yuwei Zhang, Tong Xia, Jing Han, Yu Wu, Georgios Rizos, Yang Liu, Mohammed Mosuily, Jagmohan Chauhan, and Cecilia Mascolo. Towards open respiratory acoustic foundation models: Pretraining and benchmarking, 2024.

[91] Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173, 2009.

[92] Stanley Fahn, R. L. Elton, and UPDRS Development Committee. Unified Parkinson's disease rating scale. In Stanley Fahn, C. D. Marsden, D. B. Calne, and M. Goldstein, editors, *Recent Developments in Parkinson's Disease*, volume 2, pages 153–163. Macmillan Healthcare Information, Florham Park, NJ, 1987.

[93] Amresh Bhandari and Todd Wagner. Self-reported utilization of health care services: improving measurement and accuracy. *Medical Care Research and Review*, 63(2):217–235, 2006.

[94] Alaa Althubaiti. Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of Multidisciplinary Healthcare*, 9:211–217, 2016.

[95] Christopher R. Beam, Cody Kaneshiro, Jung Yun Jang, Chandra A. Reynolds, Nancy L. Pedersen, and Margaret Gatz. Differences between women and men in incidence rates of dementia and alzheimer's disease. *Journal of Alzheimer's Disease*, 64(4):1077–1083, 2018.

[96] Ingo Hertrich and Hermann Ackermann. Acoustic analysis of speech timing in huntington's disease. *Brain and language*, 47(2):182–196, 1994.

[97] Barbara Dodd. Differential diagnosis of pediatric speech sound disorder. *Current Developmental Disorders Reports*, 1(3):189–196, 2014.

[98] Sejal Bhalla, Salaar Liaqat, Robert Wu, Andrea S. Gershon, Eyal de Lara, and Alex Mariakakis. Pulmolistener: Continuous acoustic monitoring of chronic obstructive pulmonary disease in the wild. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 7(3), September 2023.

[99] Alyssa M. Lanzi, Anna K. Saylor, Davida Fromm, Houjun Liu, Brian MacWhinney, and Matthew L. Cohen. Dementiabank: Theoretical rationale, protocol, and illustrative analyses. *American Journal of Speech-Language Pathology*, 32(2):426–438, 2023.

[100] Harold Goodglass, Edith Kaplan, and Barbara Barresi. *Boston Diagnostic Aphasia Examination*. Lippincott Williams & Wilkins, Philadelphia, PA, 3rd edition, 2001.

[101] Marshal F Folstein, Susan E Folstein, and Paul R McHugh. Mini-mental state. *Journal of psychiatric research*, 12(3):189–198, 1975.

[102] Andrew Kertesz. *The Western Aphasia Battery*. Grune & Stratton, New York, 1982.

[103] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue. TIMIT acoustic-phonetic continuous speech corpus, 1993. LDC93S1.

[104] Pamela Enderby. Frenchay dysarthria assessment. *British Journal of Disorders of Communication*, 15(3):165–173, 1980.

[105] Margaret M Hoehn and Melvin D Yahr. Parkinsonism: onset, progression, and mortality. *Neurology*, 17(5):427–427, 1967.

[106] Katherine Verdolini, Clark A. Rosen, and Ryan C. Branski, editors. *Classification Manual for Voice Disorders-I*. Psychology Press, New York, 1st edition, 2006.

[107] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673, 2020. `https://github.com/facebookresearch/libri-light`.

[108] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[109] Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio. In *Interspeech 2021*, pages 3670–3674, 2021.

[110] Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online, August 2021. Association for Computational Linguistics.

[111] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation, 2024.

[112] Chandan KA Reddy, Ebrahim Beyrami, Jamie Pool, Ross Cutler, Sriram Srinivasan, and Johannes Gehrke. A scalable noisy speech dataset and online subjective test framework. *Proc. Interspeech 2019*, pages 1816–1820, 2019.

[113] James Traer and Josh H. McDermott. Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences*, 113(48):E7856–E7865, 2016.

[114] Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, Francesco Paissan, Davide Borra, Salah Zaiem, Zeyu Zhao, Shucong Zhang, Georgios Karakasidis, Sung-Lin Yeh, Pierre Champion, Aku Rouhe, Rudolf Braun, Florian Mai, Juan Zuluaga-Gomez, Seyed Mahed Mousavi, Andreas Nautsch, Ha Nguyen, Xuechen Liu, Sangeet Sagar, Jarod Duret, Salima Mdhaffar, Gaëlle Laperrière, Mickael Rouvier, Renato De Mori, and Yannick Estève. Open-source conversational ai with speechbrain 1.0. *Journal of Machine Learning Research*, 25(333), 2024.

[115] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624.

[116] Yoshihiko Ozaki, Shuhei Watanabe, and Toshihiko Yanase. OptunaHub: A platform for black-box optimization. *arXiv preprint arXiv:2510.02798*, 2025.

Appendix for SpeechDx

A Dataset Details

This section provides a comprehensive description of the datasets utilized within the SpeechDx benchmark. The datasets are categorized according to the specific stage of speech production they represent. For each dataset, we describe the recruitment procedure, study protocol, and labeling process. Their access methods and licenses are listed in Table 2.

A.1 Conceptualization

Extended Distress Analysis Interview Corpus, Wizard-of-Oz (EDAIC-WOZ) [65] was developed at the Institute for Creative Technologies at the University of Southern California. A total of 275 participants (170 male, 105 female) drawn from the general population and American military veterans completed semi-structured clinical interviews with a virtually animated interviewer covering topics related to psychological distress, daily functioning, and mood, lasting approximately 16 minutes. For the 163 training and 56 validation participants, the interviewer was controlled by a human operator via a Wizard-of-Oz paradigm, while the 56 test participants were interviewed by a fully autonomous system, introducing a distributional shift relative to the training data. Prior to each interview, participants completed the Patient Health Questionnaire (PHQ-8) for depression severity [91]; scores of 10 or above indicate a depressive disorder, yielding binary depression labels.

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [66] comprises 24 professional actors (12 female, 12 male) narrating two lexically matched statements (*"Kids are talking by the door"* and *"Dogs are sitting by the door"*) in a neutral North American English accent while expressing one of eight emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. Each emotion (except neutral) was produced at two intensity levels (normal and strong), yielding 60 recordings per actor. The speech subset is sampled at 48 kHz. Each recording was rated 10 times for emotional validity, intensity, and authenticity by 247 untrained participants with high inter-rater reliability. The dataset is fully balanced across actors, emotions, and intensity levels.

Interactive Emotional Dyadic Motion Capture (IEMOCAP) [67] was collected at the Speech Analysis and Interpretation Laboratory at the University of Southern California. The corpus contains approximately 12 hours of audiovisual recordings from 10 actors (5 male, 5 female) across five dyadic sessions, each with a unique male-female actor pair. Within each session, actors performed both scripted emotional dialogues and improvised hypothetical scenarios designed to elicit a range of emotional states. Utterances were segmented and annotated by trained raters using categorical emotion labels and dimensional ratings (valence, activation, dominance). The original annotations include neutral, happiness, sadness, anger, surprise, fear, disgust, frustration, excited, and other emotional states. Following standard practice [67], we use a four-class subset in which *excited* is merged with *happy*, *frustration* with *anger*, and *fear*, *disgust*, *surprise*, and *other* are excluded due to their limited representation.

A.2 Formulation

DementiaBank (ADReSS-M subset [68]) was collected as part of the Alzheimer and Related Dementias Study at the University of Pittsburgh. Participants completed the Cookie Theft Picture Description Task from the Boston Diagnostic Aphasia Examination [100], producing spontaneous speech descriptions of a kitchen scene. Participants also completed the Mini-Mental State

Table 2: Access and licensing information for each dataset in SpeechDx.

Category	Dataset	Link	Availability	License
Conceptualization	EDAIC-WOZ [65]	https://dcapswoz.ict.usc.edu/	Available on request	DAIC-WOZ EULA
	RAVDESS [66]	https://zenodo.org/records/1188976	Open access	CC BY-NC-SA 4.0
	IEMOCAP [67]	https://sail.usc.edu/iemocap/iemocap_release.htm	Available on request	Custom
Formulation	DementiaBank [99, 68]	https://talkbank.org/dementia/	Available on request	TalkBank CC BY-NC-SA 3.0
	AphasiaBank [69]	https://talkbank.org/aphasia/	Available on request	TalkBank CC BY-NC-SA 3.0
Articulation (Neuromuscular)	TORGO [70]	https://www.cs.toronto.edu/~complingweb/data/TORGO/torgo.html	Open access	Custom
	UASpeech [71]	https://speechtechnology.web.illinois.edu/uaspeech/	Available on request	Custom
	MDVR-KCL [72]	https://zenodo.org/records/2867216	Open access	CC BY 4.0
	KSoF-C [17]	https://zenodo.org/records/7258757	Available on request	EULA KSoF Challenge
Articulation (Phonatory / Respiratory)	COVID-19 Sounds [74]	https://covid-19-sounds.org/en/blog/neurips_dataset	Available on request	Custom
	Coswara [75]	https://github.com/iiscleap/Coswara-Data	Open access	CC BY 4.0
	AVFAD [76]	https://acsa.web.ua.pt/AVFAD.htm	Available on request	Custom

Examination (MMSE) [101] to assess their level of cognitive impairment. We use the ADReSS-M challenge subset [68], released for the ICASSP 2023 Signal Processing Grand Challenge on multilingual Alzheimer's dementia recognition. The subset is statistically matched to mitigate common evaluation biases including repeated recordings from the same participant, audio quality variation, and demographic imbalance. The training partition consists of English-language recordings, and the test partition consists of Greek-language recordings from a separate cohort, making this the only cross-lingual evaluation in the benchmark.

AphasiaBank [69] is a publicly available database of structured discourse samples from individuals with post-stroke aphasia and neurologically healthy controls, hosted by the TalkBank consortium. We use a subset drawn from five sites — Adler, Kansas, Kurland, SCALE, and Wright — retaining first-session recordings only, yielding 206 participants (119 persons with aphasia, 87 controls). Persons with aphasia ranged in age from 36 to 91 years and spanned mild to severe impairment as measured by the Western Aphasia Battery Aphasia Quotient [102] (mean 67.6 ± 22.1). Aphasia types included Anomic, Broca's, Conduction, Wernicke's, and several transcortical variants. Each participant completed the AphasiaBank Discourse Protocol, from which we extracted four discourse tasks: (1) Cinderella story retell, (2) window picture sequence, (3) umbrella picture sequence, and (4) cat picture sequence.

A.3 Articulation (Neuromuscular)

TORGO [70] contains approximately 8 hours of English speech from 15 speakers: 8 with dysarthria (5 male, 3 female) resulting from cerebral palsy or ALS, and 7 age- and gender-matched healthy controls (4 male, 3 female). Speech was recorded using a head-mounted microphone and a directional microphone, but we only use the head-mounted channel. Speech material was drawn from multiple sources including the TIMIT database [103], lists of identified phonetic contrasts, and standardized speech intelligibility assessments. Each dysarthric speaker was assessed by a speech-language pathologist using the Frenchay Dysarthria Assessment [104].

UASpeech [71] contains isolated word recordings from 15 speakers with cerebral palsy and 13 age-matched healthy controls. Each participant uttered 765 isolated words: 300 uncommon words, 3 repetitions of 100 common words, and 165 digits, radio alphabet letters, and computer commands. Audio was recorded using an 8-microphone array, but we use the single close-talking microphone channel. Intelligibility was assessed by five listeners who transcribed the recordings word by word; each speaker's intelligibility score was computed as the average percentage of words correctly transcribed. Speakers were subsequently grouped into four intelligibility categories: very low, low, mid, and high. The distribution of dysarthric speakers across these categories is skewed toward lower intelligibility, with the very low and low groups together accounting for the majority.

MDVR-KCL [72] was released by King's College London and contains speech recordings from 37 participants: 16 with Parkinson's disease and 21 healthy controls. Recordings were made at 44.1 kHz in a phone-call scenario where participants held the recording device to their preferred ear. Participants read two passages: the fable *North Wind and the Sun* and an excerpt from *Computer Applications in Geography*. Each participant was assessed along three clinically validated scales by experts: the Hoehn and Yahr (H&Y) scale [105], Unified Parkinson's Disease Rating Scale (UPDRS) Part II Item 5 (activities of daily living and speech), and UPDRS Part III Item 18 (motor examination and speech) [92].

KSoF-C [17] is a subset of the Kassel State of Fluency (KSoF) dataset [73], a German-language stuttering corpus recorded during therapy sessions. The full KSoF corpus contains 5,597 three-second clips extracted from 214 recordings of 37 persons who stutter (28 male, 9 female). A distinctive property of this corpus is that all speakers had undergone fluency-shaping therapy, making it the only public dataset to include post-therapy stuttered speech; therefore, speakers exhibit both dysfluent events and therapy-specific speech modifications alongside fluent speech. Clips were annotated with one of six labels by three trained annotators: blocks, prolongations, sound repetitions, word repetitions, interjections, and speech modifications. KSoF-C is the subset released for the ACM Multimedia 2022 ComParE challenge [17]; we use this split with its predefined train, validation, and test partitions.

A.4 Articulation (Phonatory / Respiratory)

COVID-19 Sounds [74] was collected by the University of Cambridge via a smartphone and web application. We use the speech recordings, in which participants read the sentence *"I hope my data can help to manage the virus pandemic"* three times in their native language. The full corpus contains 53,449 samples from 36,116 participants, totaling over 552 hours of audio. Along with audio, the participants self-reported their COVID-19 status and respiratory symptoms. The dataset comes with two curated subsets widely used for benchmarking: one balanced for respiratory symptoms and the other balanced for COVID-19 status, each screened for recording quality and released as controlled English-language partitions.

Coswara [75] was collected by the Indian Institute of Science via Android and Web applications. We use the normal-pace counting speech recordings, in which participants count from 1 to 20 in English. Participants self-reported their symptoms and COVID-19 status by selecting one of three categories: negative, positive, or recovered. The control group is heterogeneous, comprising completely healthy individuals alongside those with respiratory ailments and COVID-like symptoms, reflecting realistic variability in population-level screening settings.

Advanced Voice Function Assessment Databases (AVFAD) [76] contains Portuguese recordings from 709 individuals: 346 with clinically diagnosed vocal pathology and 363 without vocal alterations. The pathological group encompasses 26 distinct diagnoses, the most prevalent being vocal fold nodules, polyps, cysts, and Reinke's edema. All diagnoses were registered according to the Classification Manual of Voice Disorders-I [106].

B Task Details

This section provides a detailed description of all 27 tasks formulated in the SpeechDx benchmark. The tasks are again organized by the stage of speech production. For each task, we describe the prediction target, the dataset used, and any relevant details regarding label construction or subset selection. The demographic statistics are summarized in Table 3.

B.1 Conceptualization

T1. This task aims to classify speakers in the EDAIC-WOZ dataset [65] as either depressed or healthy. The EDAIC-WOZ dataset contains clinical interview recordings of participants scored on the PHQ-8 depression scale [91], where speakers with a score of 10 or above are labeled as depressed.

Table 3: Demographic data and label distributions for each task in SpeechDx. Label distributions are reported as healthy % / disease % for binary classification tasks and as [range of label], mean $\pm$ standard deviation for regression tasks. Details are directly reported from publications and documentation.

Category	Dataset	ID	Country	Age	Sex	Label Distribution
Conceptualization	EDAIC-WOZ [65]	T1	USA	N/A	M: 61.8% F: 38.2%	24.0% / 76.0%
		T2	USA	N/A	M: 61.8% F: 38.2%	$[0, 23],\ \mu = 6.94 \pm 6.11$
	RAVDESS [66]	T3	Canada	N/A	M: 50% F: 50%	Calm: 13.3%; Happy: 13.3%; Sad: 13.3%; Angry: 13.3%; Fearful: 13.3%; Disgust: 13.3%; Surprised: 13.3%; Neutral: 6.7%
		T4	Canada	N/A	M: 50% F: 50%	53.3% / 46.7%
	IEMOCAP [67]	T5	USA	N/A	M: 50.8% F: 49.2%	Neutral: 23.1%; Anger / Frustrated: 40.0%; Sad: 14.7%; Happy / Excited: 22.2%
		T6	USA	N/A	M: 50.8% F: 49.2%	54.7% / 45.3%
Formulation	Dementia-Bank [99]	T7	USA; Greece	$\mu = 68.1 \pm 7.0$	M: 33.0% F: 67.0%	50.9% / 49.1%
		T8	USA; Greece	$\mu = 68.0 \pm 7.0$	M: 32.8% F: 67.2%	$[3, 30],\ \mu = 23.58 \pm 6.61$
	Aphasia-Bank [69]	T9	USA	$\mu = 62.4 \pm 13.6$	M: 55.4% F: 44.2%	72.0% / 28.0%
Articulation (neuromuscular)	TORGO [70]	T10	Canada	N/A	M: 58.1% F: 41.9%	33.8% / 66.2%
		T11	Canada	N/A	M: 55.0% F: 45.0%	$[1, 3],\ \mu = 1.84 \pm 0.94$
	UASpeech [71]	T12	USA	N/A	M: 71.4% F: 28.6%	53.6% / 46.4%
	MDVR-KCL [72]	T13	UK	N/A	N/A	42.5% / 57.5%
		T14	UK	N/A	N/A	$[0, 3],\ \mu = 0.34 \pm 0.65$
		T15	UK	N/A	N/A	$[0, 3],\ \mu = 0.40 \pm 0.70$
		T16	UK	N/A	N/A	$[0, 4],\ \mu = 1.10 \pm 1.36$
	KSoF-C [73]	T17	Germany	N/A	M: 86.0% F: 14.0%	74.1% / 25.9%
		T18	Germany	N/A	M: 86.2% F: 13.8%	Block: 20.7%; Prolongation: 12.0%; Sound rep.: 14.8%; Word rep.: 3.9%; Modified: 24.4%; Interjection: 13.0%; No dysfluency: 24.7%; Garbage: 3.1%
Articulation (phonatory / respiratory)	COVID-19 Sounds [74]	T19	Intl.	$\mu = 39.4 \pm 14.1$	M: 46.4% F: 52.3% Unk.: 1.2%	48.7% / 51.3%
		T20	Intl.	$\mu = 39.2 \pm 15.2$	M: 61.6% F: 37.2% Unk.: 1.2%	52.5% / 47.5%
		T21	Intl.	$\mu = 40.0 \pm 13.7$	M: 50.9% F: 48.4% Unk.: 0.7%	49.4% / 50.6%
		T22	Intl.	$\mu = 41.3 \pm 14.4$	M: 56.0% F: 42.9% Unk.: 1.1%	19.6% / 80.4%
		T23	Intl.	$\mu = 36.9 \pm 14.1$	M: 56.3% F: 42.6% Unk.: 1.2%	Dry cough: 49.2%; Wet cough: 23.0%; Fever: 8.1%; Headache: 22.9%; Muscle ache: 14.2%; Dizziness: 6.7%; Sore throat: 27.8%; Short breath: 13.0%; Tightness: 12.8%; Runny/blocked nose: 5.3%; Smell/taste loss: 6.1%; Runny: 16.3%
	Coswara [75]	T24	India	$\mu = 35.1 \pm 14.1$	M: 69.0% F: 30.9% Unk.: 0.1%	38.6% / 61.4%
		T25	India	$\mu = 35.2 \pm 13.9$	M: 69.9% F: 30.0% Unk.: 0.1%	32.5% / 67.5%
		T26	India	$\mu = 38.6 \pm 16.0$	M: 62.7% F: 37.2% Unk.: 0.1%	Cold: 46.4%; Cough: 61.9%; Fever: 38.7%; Diarrhoea: 5.0%; Loss of smell: 16.2%; Muscular pain: 30.9%; Breathing difficulty: 20.1%; Fatigue: 36.3%; Sore throat: 28.0%; Other respiratory: 6.8%

Continued on next page

(Table 3 continued)

Category	Dataset	ID	Country	Age	Sex	Label Distribution
	AVFAD [76]	T27	Portugal	$\mu = 52.3 \pm 15.8$	M: 29.6% F: 70.4%	48.8% / 51.2%

T2. This task aims to regress the continuous PHQ-8 depression severity score for each speaker in the EDAIC-WOZ dataset [65], using the same recordings as in T1.

T3. This task aims to classify speakers in the RAVDESS dataset [66] into one of seven emotion categories: neutral, calm, happy, sad, angry, fearful, disgust, or surprised. The RAVDESS dataset contains acted speech recordings from 24 professional actors, with each emotion represented equally across speakers.

T4. This task aims to classify speakers in the RAVDESS dataset [66] as either expressing negative or non-negative emotion using the same recordings as T3. Negative emotions comprise sad, angry, fearful, and disgust. Non-negative emotions comprise neutral, calm, happy, and surprised.

T5. This task aims to classify recordings in the IEMOCAP dataset [67] into one of four emotion categories: angry, neutral, happy, or sad. The IEMOCAP dataset contains dyadic conversation recordings across five sessions, with emotion labels assigned through majority agreement among multiple annotators.

T6. This task aims to classify speakers in the IEMOCAP dataset [67] as either expressing negative or non-negative emotion using the same recordings as T5. Negative emotions comprise angry and sad. Non-negative emotions comprise neutral and happy.

B.2 Formulation

T7. This task aims to classify speakers in the ADReSS-M challenge [68] as either having Alzheimer's disease or being cognitively healthy. The ADReSS-M challenge is a subset of the DementiaBank dataset [99] containing picture description recordings from the Cookie Theft task, with binary labels derived from clinical diagnosis of Alzheimer's disease.

T8. This task aims to regress the continuous MMSE cognitive severity score for each speaker in the DementiaBank dataset, using the same picture description recordings as T7. MMSE scores range from 0 to 30, with lower scores indicating greater cognitive impairment.

T9. This task aims to classify speakers in the AphasiaBank dataset [69] as either aphasic or healthy. The AphasiaBank dataset contains standardized discourse recordings from speakers of varying etiology and severity alongside healthy controls.

B.3 Articulation (Neuromuscular)

T10. This task aims to classify speakers in the TORGO dataset [70] as either dysarthric or healthy. The TORGO dataset contains speech recordings from speakers with motor speech disorders of varying etiology and severity alongside healthy controls.

T11. This task aims to regress the dysarthria severity score for each speaker in the TORGO dataset [70], using the same recordings as T10. Severity is rated on a scale from 0 to 3, where higher scores indicate greater articulatory impairment.

T12. This task aims to classify speakers in the UASpeech dataset [71] as either dysarthric or healthy. The UASpeech dataset contains speech recordings from speakers with cerebral palsy alongside healthy controls, with dysarthria severity ranging from mild to profound.

T13. This task aims to classify speakers in the MDVR-KCL dataset [72] as either having Parkinson's disease or being healthy. The MDVR-KCL dataset contains read speech and spontaneous monologue recordings from speakers with Parkinson's disease alongside healthy controls.

T14. This task aims to regress the UPDRS Part II Item 5 score for each speaker in the MDVR-KCL dataset [72], using the same recordings as T13. UPDRS Part II Item 5 is a patient-reported measure of speech impairment in daily living, scored from 0 to 3.

T15. This task aims to regress the UPDRS Part III Item 18 score for each speaker in the MDVR-KCL dataset, using the same recordings as T13. UPDRS Part III Item 18 is a clinician-rated measure of speech quality during motor examination, scored on a scale from 0 to 3.

T16. This task aims to regress the H&Y scale score for each speaker in the MDVR-KCL dataset, using the same recordings as T13. The Hoehn and Yahr scale is a clinician-rated measure of Parkinson's disease progression scored from 0 to 4, where higher scores indicate greater motor impairment.

T17. This task aims to classify speakers in the KSoF-C dataset [73] as either disfluent or fluent. The KSoF-C dataset contains read speech recordings from speakers who stutter alongside fluent controls, with disfluency labels assigned at the utterance level.

T18. This task aims to classify the disfluency type for each utterance in the KSoF-C dataset. This is a multi-label classification task, with each utterance being assigned one or more of six disfluency categories: modified speech, blocks, sound repetitions, interjections, prolongations, and word repetitions. Two additional classes are included to account for utterances with no disfluency and those with poor audio quality or recording errors.

B.4 Articulation (Phonatory / Respiratory)

T19. This task aims to classify speakers in the COVID-19 Sounds dataset [74] as either symptomatic or healthy. The dataset contains crowdsourced speech recordings from participants who reported their symptoms at the time of participation. Those in the symptomatic group reported at least one of the following respiratory symptoms: dry cough, sore throat, wet cough, headache, muscle ache, shortness of breath, chest tightness, fever, dizziness, smell/taste loss or runny nose. This task uses an official curated subset with controlled recording quality, balanced class distribution, and primarily English recordings.

T20. The dataset and prediction target are the same as T19, but T20 is based on the full COVID-19 Sounds dataset [74] rather than the curated subset, resulting in a larger sample size with greater variation in recording quality, language, and class balance.

T21. This task aims to classify speakers in the COVID-19 Sounds dataset [74] as either COVID-19 positive or negative based on their PCR test result. This task uses an official curated subset specifically balanced for COVID-19 status, with controlled recording quality and primarily English recordings.

T22. The dataset and prediction target are the same as T21, but T22 is based on the full COVID-19 Sounds dataset [74] rather than the curated subset, resulting in a heavily imbalanced class distribution with 20% positive and 80% negative cases.

T23. This task aims to classify the respiratory symptom type for each speaker in the COVID-19 Sounds dataset [74]. As a multi-label classification task, only speakers who reported at least one symptom are included, with each speaker labeled with one or more of the following symptom categories: dry cough, sore throat, wet cough, headache, muscle ache, shortness of breath, chest tightness, fever, dizziness, smell/taste loss, runny or blocked nose.

T24. This task aims to classify speakers in the Coswara dataset [75] as either symptomatic or healthy, where the symptomatic group consists of participants who reported at least one of the following respiratory symptoms: cough, cold, fever, fatigue, muscle pain, sore throat, breathing difficulty, loss of smell, and diarrhoea.

T25. This task aims to classify speakers in the Coswara dataset [75] as either COVID-19 positive or negative, using the same recordings as T24. The COVID-19 labels are based on self-reported test results and only speakers who reported a test outcome are included.

T26. This task aims to classify the respiratory symptom type for each speaker in the Coswara dataset [75]. As a multi-label classification task, only speakers who reported at least one symptom are included, with each speaker labeled with one or more of the following symptom categories: cough, cold, fever, fatigue, muscle pain, sore throat, breathing difficulty, loss of smell, and diarrhoea.

T27. This task aims to classify speakers in the AVFAD dataset [76] as either having a vocal pathology or not. The AVFAD dataset contains multiple speech tasks, and we use the read speech recordings for this task.

Table 4: The models evaluated on SpeechDx. Pretraining dataset sizes are approximated for models trained on aggregated or pseudo-labeled datasets. FT indicates additional fine-tuning on labeled data.

Model	Objective	Input	Supervision	Pretraining Source	Pretraining Amount (Hours)	Params (M)
wav2vec 2.0 [52]	Contrastive masked prediction	Raw audio	Self-supervised + supervised FT	Libri-Light; FT on LibriSpeech	60k + 960	317
HuBERT [53]	Masked prediction over clustered units	Raw audio	Self-supervised + supervised FT	Libri-Light; FT on LibriSpeech	60k + 960	316
WavLM [54]	Masked prediction + denoising	Raw audio	Self-supervised	Libri-Light + GigaSpeech + VoxPopuli	94k	316
MMS-1B [79]	wav2vec2.0-style SSL	Raw audio	Self-supervised	Multilingual speech corpus	$\sim$500k	1000
Qwen3-TTS-Tokenizer-12Hz [80]	Neural audio codec reconstruction	Raw audio	Self-supervised	Multilingual speech corpus	$>$5M	150
Whisper Large-v3 [81]	Sequence-to-sequence speech-to-text	Spectrogram	Supervised / pseudo-supervised	Web-scale speech–text pairs	$\sim$5M	1550
AudioMAE [83]	Masked spectrogram reconstruction	Spectrogram	Self-supervised	AudioSet	$\sim$5.8k	85.6
WavJEPA-Nat [84]	Latent representation prediction (JEPA)	Raw audio	Self-supervised	AudioSet (naturalistic scenes)	$\sim$4.8k	$\sim$200
AST [85]	Audio event classification	Spectrogram	Supervised	AudioSet	$\sim$5.8k	86.6
CLAP [86]	Audio–text contrastive learning	Spectrogram	Cross-modal contrastive	LAION-Audio-630K + AudioSet	$\sim$10k	$\sim$400
emotion2vec+ Large [89]	Multi-scale emotion learning	Raw audio	Pseudo-supervised	Large-scale emotional speech	$\sim$160k	$\sim$300
OPERA-GT [90]	Generative reconstruction	Spectrogram	Self-supervised	Multi-source respiratory audio	404	21

C Model Details

This section describes the models used in our benchmark, including their training objectives, pretraining data, and the specific variants used. All models are used with publicly available pretrained weights without additional fine-tuning unless otherwise specified.

C.1 Speech Models

wav2vec 2.0 [52] is a self-supervised speech representation model trained using a contrastive objective over discretized latent units. The large (LV-60) configuration is pretrained on approximately 60k hours of Libri-Light [107] and subsequently fine-tuned on 960 hours of LibriSpeech [108] for automatic speech recognition.

HuBERT (Large) [53] is a self-supervised speech model trained via masked prediction of offline cluster assignments derived from k-means clustering of acoustic features. Unlike wav2vec 2.0, it relies on iterative refinement of cluster targets rather than contrastive learning. The large variant is pretrained on 60k hours of Libri-Light [107] and fine-tuned on LibriSpeech [108] for automatic speech recognition.

WavLM (Large) [54] extends HuBERT by incorporating a denoising objective and training on mixtures of overlapping speech to improve robustness to real-world acoustic conditions. It combines masked prediction with augmentation strategies that simulate multi-speaker environments. The large configuration is pretrained on approximately 94k hours of speech spanning Libri-Light [107], GigaSpeech [109], and VoxPopuli [110].

MMS-1B [79] is a multilingual extension of wav2vec 2.0 designed to scale speech representation learning across languages. It is trained using self-supervised learning on approximately 500k hours of speech covering over 1,400 languages. We evaluate the 1B-parameter configuration.

Qwen3-TTS-Tokenizer (12Hz) [80] is a neural audio codec that encodes speech into discrete tokens at a fixed temporal resolution (12.5 Hz) and reconstructs the waveform from these tokens. The model

is trained using a self-supervised reconstruction objective on large-scale multilingual speech data as part of the Qwen3-TTS model.

Whisper (Large-v3) [81] is a sequence-to-sequence encoder–decoder model trained to map audio to text. It is trained on approximately 5 million hours of weakly and pseudo-labeled audio–text pairs, enabling robust performance across languages and domains. The model jointly learns multiple tasks including speech recognition, translation, and language identification through task conditioning. We evaluate the Large-v3 configuration.

C.2 General Audio Models

AudioMAE [83] is a masked autoencoder that learns audio representations by reconstructing missing regions of log-mel spectrogram inputs. A large proportion of the input patches are masked during training, encouraging the model to capture global acoustic structure. The base configuration is pretrained on approximately 5.8k hours of audio from AudioSet [82].

WavJEPA-Nat [84] is trained using a joint-embedding predictive architecture that learns to predict high-level latent representations of masked audio segments rather than reconstructing the input signal directly. The model is designed to capture semantic structure while being robust to low-level variations. It is pretrained on AudioSet [82] with additional naturalistic augmentations such as noise and reverberation. We use the naturalistic base variant.

AST [85] adapts vision transformers to audio classification by operating on log-mel spectrogram patches. It leverages ImageNet-pretrained weights for initialization and is fine-tuned on AudioSet [82] for supervised audio classification. This transfer from vision to audio enables strong performance with relatively limited audio data.

CLAP [86] is a dual-encoder model trained to align audio and text representations using a contrastive objective. We evaluate a general-domain configuration trained on diverse audio–text pairs drawn from sources such as LAION-Audio [111] and AudioSet [82] with caption augmentation rather than variants specialized for music or speech.

C.3 Domain-specific Models

emotion2vec+ (Large) [89] is designed to capture emotional characteristics in speech using a multi-stage training pipeline. It first learns representations from labeled emotional speech data and then scales to large unlabeled corpora via pseudo-labeling in a teacher–student framework. The large configuration is trained on emotional speech datasets totaling approximately 160k hours, with filtering applied during training.

OPERA-GT [90] is a domain-specific model designed for respiratory sound analysis. It is trained using a self-supervised generative objective that reconstructs spectrogram representations, along with auxiliary tasks designed to capture clinically relevant acoustic features. The model is pretrained on approximately 400 hours of curated respiratory audio data. We use the generative (GT) variant.

D Implementation Details

This section describes the implementation details of the SpeechDx benchmark, including data augmentation, loss functions, class weighting, and hyperparameter optimization.

D.1 Data Augmentation

To improve robustness and partially compensate for limited dataset sizes, each training sample is augmented to produce additional training instances. Three augmentation strategies are applied jointly: (1) additive noise drawn from the Microsoft SNSD corpus [112] at a randomly sampled signal-to-noise ratio in $[0, 15]$ dB; (2) convolutive reverberation using room impulse responses from the MIT IR Survey [113]; and (3) speed perturbation uniformly sampled from $[90\%, 110\%]$. For COVID-19 Sounds, speed perturbation is restricted to $[95\%, 105\%]$ to preserve the integrity of breathing patterns. Three augmented versions are produced per training sample for most datasets; this is increased to five for MDVR-KCL given its small sample count, and reduced to one for COVID-19 Sounds, where the full dataset is already used in training. Augmentation is applied exclusively to training data;

validation and test sets are evaluated on clean audio only. All augmentation is implemented using SpeechBrain [114, 115].

D.2 Training Setup

All binary and multi-label classification tasks are trained with binary cross-entropy loss, multi-class tasks with categorical cross-entropy loss, and regression tasks with weighted mean squared error (MSE). To mitigate the class imbalance that is prevalent across clinical datasets, all tasks use inverse-frequency weighting. For binary and multi-label tasks, positive class weights are computed per label as the ratio of negative to positive samples. For multi-class tasks, per-class weights are computed analogously. For regression tasks, sample weights are assigned via binning. Clinically established severity thresholds directly define the bins when available: MMSE scores are binned as severe (≤ 9), moderate (10–18), mild (19–23), and normal (≥ 24) [101]; and PHQ-8 scores are binned as $[0\text{–}4]$, $[5\text{–}9]$, $[10\text{–}14]$, $[15\text{–}19]$, and $[20+]$ [91]. For tasks with discrete ordinal severity ratings (T11 and T14–T16), bin edges are derived from midpoints between consecutive unique values. Each sample is assigned a weight equal to the inverse frequency of its bin, normalized so that the mean per-sample weight equals 1 to maintain a stable loss scale across tasks.

D.3 Hyperparameter Optimization

For each model-task combination, hyperparameters are selected via Optuna [116] over five trials to optimize validation loss. The learning rate is sampled log-uniformly over $[10^{-4}, 10^{-3}]$ and decayed linearly to one-tenth of its initial value over the training run. Weight decay (L2 regularization) is sampled log-uniformly over $[0.01, 0.1]$. All models are trained with a batch size of 16 in 32-bit floating-point precision for up to 50 epochs. Early stopping is triggered after 5 consecutive epochs without improvement in validation loss; the first 4 epochs are excluded from the early stopping criterion to allow sufficient warm-up. All experiments use a global random seed.

E Data Efficiency Analysis

Clinical speech datasets are typically small, and label acquisition is costly. To characterize how different encoders perform under data scarcity, we train linear probes using 12.5%, 25%, 50%, and 100% of the available training data. This analysis is conducted for Qwen3-TTS-Tokenizer [80], WavLM [54], and Whisper [81], which were identified as the top-performing models in Section 5.1. For each data regime, training subsets are constructed using the same stratification protocol as the main experiments. This ensures that the label distribution is preserved across all data regimes. All models use identical subsampled splits at each fraction, and all probes are evaluated on the same test set. All per-task results are reported in Table 5, while Figure 3 summarizes AUC trajectories grouped by speech production stage.

Qwen3 emerges as the most data-efficient model, leading 11 out of 27 tasks when trained on the least amount of data. Whisper exhibits competitive performance at higher amounts of training data but demonstrates reduced effectiveness at the lowest fraction. WavLM exhibits substantial performance variability across tasks and data regimes; while it achieves strong performance on specific tasks at low amounts of training data (e.g., AUC: 0.54 for T1 at 12.5%, AUC: 0.85 for T12 at 12.5%), it also demonstrates severe performance degradation on others (e.g., AUC: 0.33 for T7 at 12.5%). This variability precludes reliable deployment of WavLM in low-data regimes without task-specific validation.

At the task level, several conditions demonstrate strong performance with minimal training data. Aphasia detection (T9) achieves robust performance even at 12.5% of training data, with Qwen3 reaching an AUC of 0.90 and improving to 0.97 at 100%, indicating that aphasia-relevant speech patterns can be captured with limited supervision. Similarly, dysarthria detection (T10, T12), disfluency classification (T18), and multiple respiratory symptom tasks (T19, T20, T21, T23, T25, T26) exhibit relatively flat learning curves, with most models approaching within 0.05-0.10 AUC of their full-data performance using only 25% of the training set.

In contrast, Parkinson's detection tasks (T13, T14, T16) and Alzheimer's detection (T7) demonstrate the steepest data requirements. For T13, the AUC of Whisper improves from 0.61 to 0.73, while WavLM exhibits a more pronounced improvement from an AUC of 0.56 to 0.81. It is important to

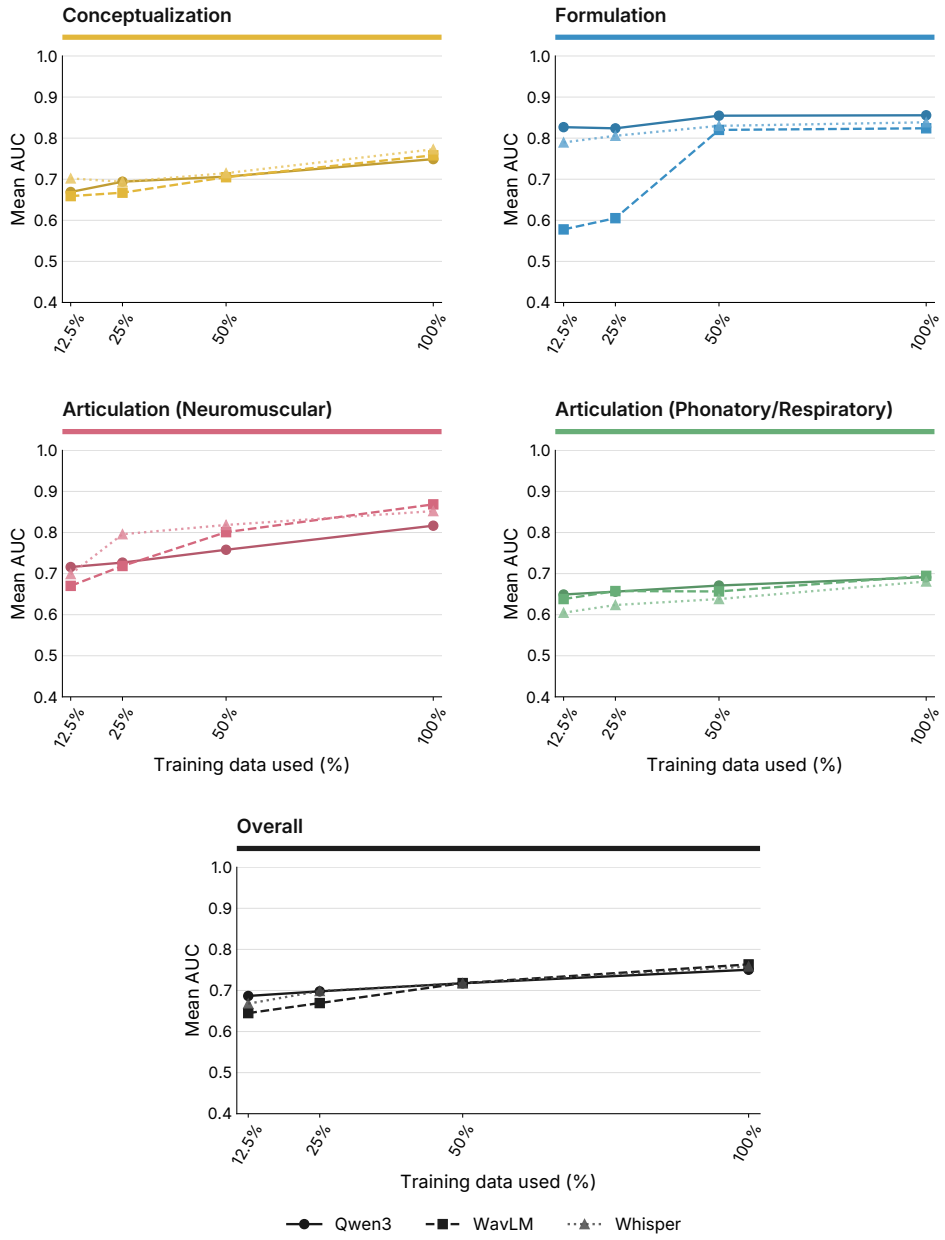

Figure 3: The data efficiency of Qwen3-TTS-Tokenizer [80], WavLM [54], and Whisper [81] with various amounts of training data.

note that these proportions of training data reflect different absolute sample sizes across tasks: 12.5% of training data corresponds to as few as 4 training samples for small datasets like MDVR-KCL (30 total subjects) and as many as 3,156 samples for larger datasets like COVID-19 Sounds (25,252 total samples), which partially accounts for the observed task-level variability in data efficiency.

These results expose substantial heterogeneity in data efficiency across both tasks and models. While the majority of tasks plateau by 25-50% of available training data, specific conditions consistently require larger training sets across all evaluated encoders. Significant instability under data scarcity indicates that current self-supervised speech encoders lack robustness when training supervision is limited. Overall, data efficiency depends critically on the interaction between encoder architecture, pretraining strategy, and the specific acoustic manifestations of each clinical condition.

Table 5: Data-efficiency results. Per-task performance for Qwen3, WavLM, and Whisper when trained on 12.5%, 25%, 50%, and 100% of the available training data. Each cell reports the metric value with the bootstrapped 95% confidence interval below. The best of the three models for each task and training fraction is shown in bold.

Task	Metric	12.5%			25%			50%			100%		
		Qwen3	WavLM	Whisper	Qwen3	WavLM	Whisper	Qwen3	WavLM	Whisper	Qwen3	WavLM	Whisper
Conceptualization													
T1 (Depression detection)	AUC	0.43 (0.27, 0.59)	**0.54** (0.36, 0.70)	0.47 (0.29, 0.63)	0.46 (0.29, 0.63)	**0.49** (0.32, 0.65)	0.38 (0.20, 0.55)	0.44 (0.26, 0.61)	**0.47** (0.30, 0.63)	0.39 (0.21, 0.56)	0.46 (0.29, 0.63)	**0.57** (0.40, 0.73)	0.55 (0.37, 0.72)
T2 (Depression severity)	MAE	5.34 (4.34, 6.43)	7.09 (5.67, 8.73)	**5.26** (4.31, 6.23)	**5.25** (4.32, 6.23)	6.62 (5.28, 8.17)	5.25 (4.32, 6.16)	**5.42** (4.34, 6.59)	5.84 (4.74, 7.10)	5.48 (4.45, 6.59)	**5.31** (4.30, 6.38)	5.37 (4.40, 6.37)	5.34 (4.35, 6.35)
T3 (Emotion classification)	AUC	0.74 (0.70, 0.77)	0.72 (0.70, 0.74)	**0.81** (0.80, 0.83)	0.78 (0.74, 0.83)	0.76 (0.72, 0.80)	**0.82** (0.80, 0.84)	0.84 (0.82, 0.86)	0.81 (0.79, 0.84)	**0.85** (0.83, 0.88)	**0.89** (0.86, 0.91)	0.86 (0.84, 0.88)	0.87 (0.85, 0.90)
T4 (Binary emotion classification)	AUC	0.62 (0.58, 0.67)	0.59 (0.48, 0.70)	**0.72** (0.68, 0.75)	0.68 (0.60, 0.75)	0.65 (0.54, 0.76)	**0.76** (0.72, 0.79)	0.66 (0.64, 0.68)	0.70 (0.61, 0.80)	**0.76** (0.73, 0.79)	0.78 (0.75, 0.81)	0.76 (0.75, 0.78)	**0.82** (0.79, 0.84)
T5 (Emotion classification)	AUC	0.81 (0.78, 0.84)	0.78 (0.76, 0.81)	**0.82** (0.80, 0.84)	0.81 (0.78, 0.84)	0.78 (0.76, 0.81)	**0.82** (0.80, 0.84)	0.82 (0.80, 0.85)	0.81 (0.78, 0.84)	**0.84** (0.81, 0.87)	0.83 (0.81, 0.86)	0.83 (0.81, 0.85)	**0.86** (0.85, 0.88)
T6 (Binary emotion classification)	AUC	**0.75** (0.70, 0.79)	0.66 (0.62, 0.70)	0.69 (0.68, 0.71)	**0.75** (0.70, 0.79)	0.66 (0.62, 0.70)	0.69 (0.68, 0.71)	**0.77** (0.74, 0.80)	0.73 (0.70, 0.76)	0.73 (0.70, 0.76)	**0.78** (0.75, 0.81)	0.76 (0.73, 0.80)	0.76 (0.73, 0.79)
Formulation													
T7 (Alzheimer's detection)	AUC	0.75 (0.59, 0.89)	0.33 (0.17, 0.50)	**0.76** (0.61, 0.90)	**0.75** (0.59, 0.89)	0.31 (0.16, 0.49)	0.73 (0.56, 0.87)	**0.76** (0.59, 0.90)	0.69 (0.53, 0.84)	0.75 (0.59, 0.89)	0.74 (0.58, 0.88)	0.69 (0.53, 0.83)	**0.75** (0.59, 0.89)
T8 (Alzheimer's severity)	MAE	**13.44** (11.68, 15.04)	21.92 (20.31, 23.37)	18.87 (17.21, 20.37)	**12.02** (10.35, 13.56)	18.91 (17.25, 20.44)	13.89 (12.15, 15.42)	11.08 (9.44, 12.68)	13.61 (11.93, 15.16)	**9.67** (8.29, 10.98)	10.75 (9.25, 12.20)	**8.58** (7.23, 9.87)	9.97 (8.60, 11.33)
T9 (Aphasia detection)	AUC	**0.90** (0.81, 0.98)	0.83 (0.69, 0.94)	0.82 (0.68, 0.93)	**0.90** (0.79, 0.99)	**0.90** (0.79, 0.99)	0.89 (0.78, 0.97)	**0.95** (0.89, 0.99)	0.95 (0.89, 0.99)	0.91 (0.82, 0.98)	**0.97** (0.92, 1.00)	0.96 (0.91, 0.99)	0.92 (0.84, 0.98)
Articulation (Neuromuscular)													
T10 (Dysarthria detection)	AUC	**0.79** (0.67, 0.91)	0.59 (0.44, 0.75)	0.61 (0.38, 0.84)	0.68 (0.37, 0.99)	0.66 (0.48, 0.85)	**0.84** (0.71, 0.97)	0.75 (0.44, 1.06)	0.75 (0.48, 1.03)	**0.87** (0.75, 1.00)	0.87 (0.77, 0.97)	0.88 (0.76, 1.00)	**0.91** (0.84, 0.98)
T11 (Dysarthria severity)	MAE	**0.56** (0.36, 0.76)	0.57 (0.35, 0.78)	0.61 (0.32, 0.89)	0.49 (0.33, 0.65)	0.54 (0.33, 0.74)	**0.37** (0.27, 0.48)	0.61 (0.31, 0.92)	0.61 (0.38, 0.84)	**0.47** (0.29, 0.66)	0.54 (0.33, 0.75)	0.63 (0.37, 0.90)	**0.43** (0.27, 0.59)
T12 (Dysarthria detection)	AUC	**0.93** (0.87, 0.99)	0.85 (0.63, 1.06)	0.90 (0.84, 0.96)	0.92 (0.85, 0.98)	0.90 (0.80, 1.01)	**0.94** (0.87, 1.00)	0.92 (0.81, 1.03)	0.91 (0.76, 1.05)	**0.95** (0.89, 1.01)	0.95 (0.90, 0.99)	0.96 (0.91, 1.01)	**0.97** (0.92, 1.01)
T13 (Parkinson's detection)	AUC	0.55 (0.37, 0.74)	0.56 (0.40, 0.72)	**0.61** (0.50, 0.71)	0.60 (0.41, 0.80)	0.56 (0.45, 0.68)	**0.73** (0.66, 0.81)	0.61 (0.52, 0.70)	**0.74** (0.61, 0.86)	0.72 (0.63, 0.80)	0.67 (0.51, 0.83)	**0.81** (0.75, 0.88)	0.73 (0.65, 0.81)
T14 (Parkinson's severity)	MAE	0.45 (0.39, 0.50)	**0.44** (0.39, 0.49)	0.46 (0.43, 0.49)	**0.43** (0.37, 0.49)	0.44 (0.40, 0.48)	0.46 (0.43, 0.48)	**0.44** (0.40, 0.48)	0.45 (0.42, 0.47)	0.45 (0.42, 0.48)	**0.42** (0.36, 0.49)	0.44 (0.42, 0.47)	0.43 (0.39, 0.46)
T15 (Parkinson's severity)	MAE	0.45 (0.39, 0.51)	**0.44** (0.38, 0.50)	0.46 (0.42, 0.50)	0.45 (0.39, 0.52)	**0.44** (0.37, 0.50)	0.46 (0.43, 0.50)	**0.44** (0.35, 0.52)	0.44 (0.38, 0.50)	0.46 (0.42, 0.50)	**0.42** (0.33, 0.51)	0.44 (0.38, 0.49)	0.44 (0.39, 0.49)
T16 (Parkinson's severity)	MAE	0.45 (0.41, 0.50)	**0.45** (0.40, 0.50)	0.47 (0.45, 0.50)	0.46 (0.41, 0.51)	**0.45** (0.40, 0.50)	0.47 (0.44, 0.50)	0.46 (0.40, 0.52)	**0.46** (0.42, 0.49)	0.46 (0.43, 0.49)	**0.41** (0.34, 0.48)	0.45 (0.41, 0.48)	0.44 (0.41, 0.46)
T17 (Disfluency detection)	AUC	0.68 (0.56, 0.80)	0.71 (0.65, 0.78)	**0.74** (0.69, 0.78)	0.76 (0.66, 0.87)	0.77 (0.68, 0.85)	**0.77** (0.71, 0.83)	0.77 (0.72, 0.83)	**0.82** (0.79, 0.86)	0.79 (0.75, 0.83)	0.81 (0.75, 0.86)	**0.86** (0.83, 0.88)	0.85 (0.81, 0.89)
T18 (Disfluency classification)	AUC	0.62 (0.60, 0.65)	**0.64** (0.61, 0.67)	0.64 (0.57, 0.70)	0.67 (0.63, 0.71)	0.70 (0.66, 0.74)	**0.70** (0.68, 0.72)	0.74 (0.71, 0.76)	**0.78** (0.77, 0.79)	0.76 (0.75, 0.78)	0.78 (0.78, 0.79)	**0.84** (0.82, 0.85)	0.81 (0.78, 0.83)
Articulation (Phonatory / Respiratory)													
T19 (Respiratory symptom detection)	AUC	**0.58** (0.55, 0.61)	0.57 (0.54, 0.61)	0.55 (0.51, 0.58)	0.62 (0.58, 0.65)	**0.63** (0.60, 0.66)	0.57 (0.54, 0.61)	**0.64** (0.60, 0.68)	0.62 (0.59, 0.66)	0.58 (0.55, 0.62)	**0.69** (0.66, 0.72)	0.68 (0.65, 0.71)	0.68 (0.65, 0.71)
T20 (Respiratory symptom detection)	AUC	**0.59** (0.58, 0.60)	0.58 (0.56, 0.59)	0.56 (0.55, 0.58)	0.59 (0.58, 0.60)	0.59 (0.58, 0.61)	**0.60** (0.58, 0.61)	0.61 (0.60, 0.62)	**0.62** (0.60, 0.63)	0.62 (0.60, 0.63)	0.64 (0.62, 0.65)	0.65 (0.64, 0.66)	**0.65** (0.64, 0.66)
T21 (COVID-19 detection)	AUC	**0.64** (0.56, 0.72)	0.54 (0.46, 0.62)	0.49 (0.41, 0.57)	**0.60** (0.52, 0.69)	0.56 (0.48, 0.64)	0.50 (0.41, 0.58)	**0.59** (0.50, 0.67)	0.49 (0.41, 0.56)	0.49 (0.40, 0.57)	0.61 (0.52, 0.69)	**0.65** (0.56, 0.72)	0.61 (0.53, 0.69)
T22 (COVID-19 detection)	AUC	0.63 (0.57, 0.67)	**0.64** (0.58, 0.69)	0.60 (0.55, 0.65)	0.64 (0.59, 0.69)	**0.66** (0.61, 0.71)	0.65 (0.59, 0.69)	0.61 (0.56, 0.66)	**0.65** (0.60, 0.69)	0.62 (0.57, 0.66)	0.65 (0.59, 0.70)	0.68 (0.63, 0.73)	**0.70** (0.65, 0.74)
T23 (Respiratory symptom classification)	AUC	**0.57** (0.57, 0.58)	0.57 (0.56, 0.58)	0.56 (0.55, 0.57)	**0.59** (0.58, 0.59)	0.58 (0.57, 0.59)	0.58 (0.57, 0.58)	0.60 (0.59, 0.60)	0.60 (0.59, 0.60)	**0.60** (0.60, 0.61)	0.60 (0.59, 0.61)	0.60 (0.59, 0.61)	**0.61** (0.61, 0.62)
T24 (Respiratory symptom detection)	AUC	0.68 (0.63, 0.72)	0.68 (0.64, 0.72)	**0.69** (0.65, 0.74)	0.69 (0.64, 0.73)	0.70 (0.66, 0.74)	**0.71** (0.67, 0.75)	0.71 (0.67, 0.76)	0.71 (0.67, 0.75)	**0.72** (0.67, 0.76)	**0.73** (0.69, 0.77)	0.72 (0.68, 0.76)	0.72 (0.68, 0.76)
T25 (COVID-19 detection)	AUC	**0.76** (0.71, 0.81)	0.75 (0.70, 0.79)	0.73 (0.68, 0.77)	0.73 (0.68, 0.78)	0.74 (0.68, 0.78)	**0.75** (0.70, 0.79)	**0.78** (0.73, 0.83)	0.74 (0.69, 0.78)	0.76 (0.71, 0.80)	**0.79** (0.74, 0.83)	0.77 (0.72, 0.81)	0.77 (0.72, 0.81)
T26 (Respiratory symptom classification)	AUC	0.55 (0.51, 0.58)	**0.55** (0.51, 0.58)	0.54 (0.51, 0.57)	0.55 (0.52, 0.59)	**0.56** (0.53, 0.60)	0.55 (0.52, 0.58)	**0.58** (0.54, 0.61)	0.57 (0.53, 0.61)	0.57 (0.53, 0.60)	**0.60** (0.57, 0.63)	0.59 (0.56, 0.63)	0.59 (0.55, 0.62)
T27 (Vocal pathology detection)	AUC	0.85 (0.80, 0.90)	**0.87** (0.83, 0.92)	0.72 (0.67, 0.77)	0.89 (0.85, 0.93)	**0.90** (0.86, 0.94)	0.72 (0.65, 0.78)	0.92 (0.87, 0.95)	**0.92** (0.87, 0.96)	0.79 (0.73, 0.85)	**0.92** (0.87, 0.95)	0.91 (0.87, 0.95)	0.81 (0.75, 0.86)

F Zero-shot Transfer Analysis

Beyond the aggregate cross-category patterns reported in Section 5.2, individual task-level transfer pairs reveal several insights into which clinical manifestations produce generalizable representations and which tasks exhibit fragmentation. Table 6 reports complete zero-shot transfer results for all model-source-target combinations.

Transfer Asymmetries Within Production Stages. Bidirectional transfer within the same production stage reveals which task representations are more generalizable. Alzheimer's $\rightarrow$ Aphasia (T7 $\rightarrow$ T9) achieves an AUC of 0.94 (HuBERT) while Aphasia $\rightarrow$ Alzheimer's (T9 $\rightarrow$ T7) reaches only 0.74 (Whisper), indicating that dementia-trained representations capture broader language disruption patterns that extend to aphasia, whereas aphasia-specific features are less informative for Alzheimer's detection. Similarly, dysarthria transfer is asymmetric: TORGO $\rightarrow$ UASpeech (T10 $\rightarrow$ T12) achieves an AUC of 0.92 (Whisper) while UASpeech $\rightarrow$ TORGO (T12 $\rightarrow$ T10) reaches only 0.76. TORGO includes speakers with more severe dysarthria (primarily cerebral palsy), while UASpeech covers a broader range of severities. These asymmetries suggest that certain clinical manifestations yield more generalizable acoustic signatures than others, though the mechanisms differ: severity in the case of dysarthria, and potentially the involvement of multiple linguistic domains in the case of dementia, versus focal language impairment in aphasia.

COVID-19 and Respiratory Task Fragmentation. COVID-19 and respiratory tasks show minimal cross-transfer. COVID-19 Sounds $\rightarrow$ Coswara achieves a maximum AUC of 0.79, within-dataset tasks (T21 $\rightarrow$ T22 and T22 $\rightarrow$ T21) achieve a maximum AUC of 0.64, and general respiratory symptoms (T19, T24) transfer to COVID-19 detection (T21, T22, T25) with AUC $\leq$ 0.69. This fragmentation reflects differences in recording protocols (cough vs. sustained phonation) or population characteristics, and that current encoders fail to extract protocol-invariant respiratory features.

Sample Size versus Feature Alignment in Emotion → Depression Transfer. Emotion-labeled data provides modest benefits for depression detection, likely through sample size rather than feature alignment. Emotion → Depression transfer (maximum AUC: 0.75) exceeds performance when trained and tested for depression itself (maximum AUC: 0.65), but the gain is small relative to the multi-fold increase in training samples (4,246 vs. 163), indicating that general affective prosody and clinical depression capture partially overlapping but distinct acoustic patterns.

Cross-category Transfer Reveals Acoustic Hierarchy. Phonatory / Respiratory → Conceptualization (maximum AUC: 0.83) and Phonatory / Respiratory → Formulation (maximum AUC: 0.88) succeed, while reverse transfers fail (AUC $\leq$ 0.60). Conceptualization tasks produce the weakest cross-category transfer as sources (AUC $\leq$ 0.67). This asymmetry suggests that low-level acoustic features, such as voice quality and phonation characteristics, provide useful priors for higher-level cognitive-linguistic tasks, raising the possibility that multi-stage screening pipelines could leverage simpler phonatory tasks as initial filters before deploying more specialized models for cognitive assessment.

Table 6: Zero-shot transfer performance within and across categories, reported as AUC with 95% confidence intervals.

Source Task ID	Source Task	Target Task ID	Target Task	Qwen3	WavLM	AST	AudioMAE	CLAP	emotion2vec+	HuBERT	MMS	OPERA-GT	wav2vec 2.0	WavJEPA	Whisper
Conceptualization															
T4	Emotion Classification	T6	Emotion Classification	0.49 (0.48, 0.50)	0.49 (0.46, 0.52)	0.57 (0.54, 0.59)	0.54 (0.52, 0.56)	0.54 (0.52, 0.56)	**0.86** (0.84, 0.87)	0.44 (0.41, 0.47)	0.49 (0.47, 0.51)	0.56 (0.54, 0.57)	0.37 (0.34, 0.40)	0.55 (0.54, 0.57)	0.58 (0.56, 0.60)
T6	Emotion Classification	T4	Emotion Classification	0.56 (0.52, 0.60)	0.54 (0.48, 0.60)	0.58 (0.55, 0.62)	0.65 (0.61, 0.69)	0.56 (0.53, 0.59)	**0.82** (0.79, 0.84)	0.59 (0.54, 0.63)	0.44 (0.40, 0.48)	0.51 (0.47, 0.55)	0.41 (0.38, 0.43)	0.62 (0.59, 0.65)	0.71 (0.67, 0.75)
T6	Emotion Classification	T1	Depression Detection	0.58 (0.50, 0.65)	0.50 (0.43, 0.58)	0.41 (0.34, 0.48)	0.44 (0.36, 0.52)	0.43 (0.36, 0.52)	**0.75** (0.68, 0.81)	0.56 (0.48, 0.64)	0.43 (0.35, 0.51)	0.45 (0.37, 0.53)	0.54 (0.46, 0.63)	0.58 (0.50, 0.65)	0.50 (0.43, 0.57)
T4	Emotion Classification	T1	Depression Detection	0.38 (0.31, 0.46)	0.41 (0.33, 0.50)	0.39 (0.31, 0.46)	0.39 (0.32, 0.47)	0.42 (0.34, 0.49)	**0.74** (0.67, 0.80)	0.47 (0.39, 0.56)	0.58 (0.49, 0.66)	0.40 (0.32, 0.48)	0.46 (0.38, 0.55)	0.57 (0.49, 0.65)	0.40 (0.33, 0.47)
T1	Depression Detection	T4	Emotion Classification	0.35 (0.31, 0.38)	0.56 (0.51, 0.60)	0.43 (0.40, 0.46)	0.45 (0.42, 0.48)	0.53 (0.49, 0.56)	**0.93** (0.90, 0.95)	0.40 (0.36, 0.44)	0.44 (0.40, 0.48)	0.42 (0.38, 0.45)	0.54 (0.50, 0.58)	0.49 (0.47, 0.52)	0.49 (0.44, 0.54)
T1	Depression Detection	T6	Emotion Classification	0.58 (0.56, 0.60)	0.54 (0.51, 0.56)	0.45 (0.42, 0.47)	0.54 (0.51, 0.56)	0.52 (0.50, 0.54)	**0.82** (0.79, 0.84)	0.61 (0.58, 0.64)	0.45 (0.43, 0.47)	0.43 (0.41, 0.45)	0.55 (0.53, 0.57)	0.53 (0.50, 0.55)	0.59 (0.56, 0.62)
Formulation															
T9	Aphasia Detection	T7	Alzheimer's Detection	0.70 (0.64, 0.76)	0.74 (0.68, 0.80)	0.65 (0.58, 0.71)	0.65 (0.58, 0.72)	0.57 (0.50, 0.63)	0.48 (0.41, 0.55)	0.73 (0.67, 0.79)	0.74 (0.68, 0.80)	0.66 (0.60, 0.72)	0.69 (0.62, 0.74)	0.68 (0.62, 0.73)	**0.74** (0.68, 0.80)
T7	Alzheimer's Detection	T9	Aphasia Detection	0.82 (0.77, 0.87)	0.90 (0.86, 0.93)	0.84 (0.79, 0.89)	0.84 (0.78, 0.88)	0.67 (0.59, 0.75)	0.61 (0.56, 0.66)	**0.94** (0.92, 0.96)	0.83 (0.78, 0.88)	0.88 (0.83, 0.91)	0.78 (0.72, 0.84)	0.83 (0.77, 0.88)	0.88 (0.83, 0.92)
Articulation (Neuromuscular)															
T10	Dysarthria Detection	T17	Disfluency Detection	0.60 (0.48, 0.68)	**0.73** (0.65, 0.80)	0.59 (0.50, 0.67)	0.57 (0.45, 0.66)	0.51 (0.41, 0.64)	0.56 (0.53, 0.60)	0.71 (0.67, 0.76)	0.70 (0.62, 0.77)	0.63 (0.55, 0.70)	0.66 (0.61, 0.71)	0.70 (0.60, 0.78)	0.73 (0.66, 0.77)
T12	Dysarthria Detection	T17	Disfluency Detection	0.66 (0.58, 0.72)	0.64 (0.57, 0.70)	0.62 (0.54, 0.69)	0.60 (0.52, 0.67)	0.50 (0.40, 0.62)	0.60 (0.55, 0.64)	0.57 (0.52, 0.61)	0.57 (0.44, 0.66)	0.61 (0.51, 0.69)	0.65 (0.60, 0.69)	0.68 (0.60, 0.74)	**0.70** (0.65, 0.74)
T10	Dysarthria Detection	T12	Dysarthria Detection	0.83 (0.72, 0.92)	0.90 (0.83, 0.96)	0.78 (0.63, 0.90)	0.78 (0.62, 0.90)	0.27 (0.16, 0.39)	0.76 (0.65, 0.86)	0.91 (0.83, 0.96)	0.80 (0.68, 0.90)	0.55 (0.41, 0.68)	0.82 (0.72, 0.91)	0.52 (0.39, 0.64)	**0.92** (0.84, 0.98)
T12	Dysarthria Detection	T13	Parkinson's Detection	0.56 (0.41, 0.70)	0.67 (0.56, 0.79)	0.54 (0.37, 0.73)	**0.83** (0.70, 0.93)	0.73 (0.59, 0.85)	0.67 (0.54, 0.80)	0.67 (0.51, 0.83)	0.61 (0.50, 0.72)	0.30 (0.16, 0.47)	0.55 (0.48, 0.62)	0.79 (0.66, 0.88)	0.65 (0.56, 0.73)
T12	Dysarthria Detection	T10	Dysarthria Detection	0.62 (0.45, 0.81)	0.72 (0.54, 0.89)	0.66 (0.48, 0.84)	0.63 (0.45, 0.81)	0.62 (0.46, 0.78)	0.71 (0.60, 0.82)	0.68 (0.53, 0.84)	0.60 (0.41, 0.80)	0.51 (0.36, 0.67)	0.71 (0.59, 0.86)	0.71 (0.58, 0.86)	**0.76** (0.58, 0.93)
T10	Dysarthria Detection	T13	Parkinson's Detection	0.50 (0.34, 0.65)	0.52 (0.36, 0.68)	0.46 (0.28, 0.65)	0.65 (0.48, 0.80)	0.71 (0.54, 0.85)	0.42 (0.25, 0.60)	0.56 (0.41, 0.72)	0.56 (0.40, 0.73)	0.45 (0.28, 0.63)	0.54 (0.43, 0.65)	0.51 (0.34, 0.67)	**0.72** (0.57, 0.85)
T13	Parkinson's Detection	T12	Dysarthria Detection	0.47 (0.32, 0.62)	0.62 (0.53, 0.71)	0.71 (0.58, 0.81)	0.53 (0.39, 0.67)	**0.81** (0.75, 0.87)	0.70 (0.62, 0.77)	0.74 (0.69, 0.79)	0.26 (0.18, 0.32)	0.28 (0.17, 0.40)	0.35 (0.26, 0.44)	0.58 (0.46, 0.70)	0.48 (0.36, 0.58)
T13	Parkinson's Detection	T10	Dysarthria Detection	0.30 (0.17, 0.46)	0.48 (0.37, 0.57)	0.58 (0.38, 0.80)	0.49 (0.35, 0.65)	**0.60** (0.48, 0.72)	0.55 (0.43, 0.65)	0.56 (0.49, 0.63)	0.55 (0.44, 0.65)	0.58 (0.50, 0.67)	0.40 (0.30, 0.53)	0.47 (0.39, 0.55)	0.56 (0.47, 0.63)
T13	Parkinson's Detection	T17	Disfluency Detection	0.49 (0.40, 0.58)	0.57 (0.51, 0.64)	0.50 (0.39, 0.61)	0.41 (0.33, 0.50)	0.51 (0.44, 0.56)	0.55 (0.49, 0.60)	0.41 (0.36, 0.45)	0.57 (0.53, 0.62)	0.46 (0.39, 0.55)	**0.58** (0.52, 0.64)	0.54 (0.43, 0.63)	0.52 (0.45, 0.58)
T17	Disfluency Detection	T13	Parkinson's Detection	0.29 (0.15, 0.45)	0.50 (0.32, 0.68)	0.57 (0.42, 0.73)	0.34 (0.19, 0.49)	0.37 (0.17, 0.59)	0.47 (0.32, 0.63)	0.37 (0.27, 0.47)	**0.76** (0.60, 0.89)	0.44 (0.24, 0.65)	0.53 (0.43, 0.63)	0.58 (0.42, 0.73)	0.55 (0.41, 0.70)
T17	Disfluency Detection	T12	Dysarthria Detection	0.76 (0.65, 0.85)	0.72 (0.61, 0.83)	0.53 (0.40, 0.64)	0.72 (0.59, 0.83)	0.52 (0.38, 0.67)	0.81 (0.71, 0.89)	0.61 (0.50, 0.71)	0.54 (0.43, 0.65)	0.76 (0.65, 0.85)	0.62 (0.53, 0.70)	0.52 (0.43, 0.62)	**0.88** (0.80, 0.95)
T17	Disfluency Detection	T10	Dysarthria Detection	0.66 (0.55, 0.80)	0.73 (0.59, 0.85)	0.57 (0.46, 0.70)	0.66 (0.50, 0.83)	0.37 (0.19, 0.56)	0.69 (0.60, 0.78)	0.74 (0.61, 0.88)	0.69 (0.53, 0.84)	0.48 (0.33, 0.65)	0.62 (0.54, 0.71)	0.64 (0.55, 0.76)	**0.74** (0.61, 0.90)
Articulation (Phonatory / Respiratory)															
T24	Respiratory Symptoms	T19	Respiratory Symptoms	0.54 (0.53, 0.56)	0.55 (0.54, 0.57)	0.53 (0.52, 0.55)	0.50 (0.49, 0.51)	0.52 (0.51, 0.54)	0.58 (0.57, 0.60)	0.55 (0.54, 0.57)	**0.61** (0.60, 0.62)	0.49 (0.48, 0.50)	0.53 (0.51, 0.54)	0.54 (0.53, 0.55)	0.51 (0.50, 0.53)
T27	Vocal pathology detection	T19	Respiratory Symptoms	0.51 (0.49, 0.52)	0.58 (0.57, 0.59)	0.53 (0.51, 0.54)	0.56 (0.54, 0.57)	0.48 (0.47, 0.50)	**0.58** (0.57, 0.60)	0.57 (0.56, 0.59)	0.58 (0.56, 0.59)	0.53 (0.52, 0.55)	0.54 (0.53, 0.56)	0.56 (0.55, 0.58)	0.54 (0.53, 0.56)
T27	Vocal pathology detection	T21	COVID-19 Detection	0.46 (0.42, 0.50)	0.48 (0.45, 0.52)	0.46 (0.43, 0.50)	0.50 (0.46, 0.53)	0.50 (0.46, 0.55)	**0.57** (0.54, 0.61)	0.52 (0.48, 0.56)	0.53 (0.50, 0.57)	0.47 (0.43, 0.51)	0.47 (0.43, 0.51)	0.51 (0.48, 0.55)	0.52 (0.48, 0.56)
T27	Vocal pathology detection	T25	COVID-19 Detection	0.41 (0.38, 0.43)	0.42 (0.40, 0.45)	0.44 (0.42, 0.46)	0.38 (0.36, 0.41)	0.50 (0.47, 0.52)	0.62 (0.60, 0.64)	**0.69** (0.67, 0.71)	0.56 (0.54, 0.58)	0.49 (0.47, 0.51)	0.54 (0.52, 0.56)	0.49 (0.46, 0.51)	0.65 (0.62, 0.67)
T27	Vocal pathology detection	T24	Respiratory Symptoms	0.42 (0.40, 0.44)	0.46 (0.44, 0.48)	0.48 (0.46, 0.50)	0.42 (0.40, 0.45)	0.47 (0.45, 0.50)	0.61 (0.59, 0.62)	**0.69** (0.67, 0.70)	0.59 (0.57, 0.61)	0.50 (0.49, 0.52)	0.58 (0.56, 0.60)	0.51 (0.49, 0.53)	0.65 (0.63, 0.67)
T19	Respiratory Symptoms	T27	Vocal pathology detection	0.58 (0.55, 0.61)	0.55 (0.53, 0.58)	0.65 (0.62, 0.68)	0.66 (0.62, 0.69)	0.43 (0.40, 0.47)	0.60 (0.57, 0.62)	0.59 (0.56, 0.61)	0.56 (0.54, 0.59)	0.50 (0.47, 0.54)	0.56 (0.54, 0.59)	**0.68** (0.65, 0.71)	0.57 (0.55, 0.59)
T21	COVID-19 Detection	T27	Vocal pathology detection	0.43 (0.40, 0.45)	0.45 (0.43, 0.47)	0.46 (0.44, 0.49)	0.42 (0.40, 0.45)	0.53 (0.49, 0.57)	**0.62** (0.60, 0.65)	0.54 (0.52, 0.56)	0.53 (0.51, 0.55)	0.52 (0.49, 0.56)	0.55 (0.53, 0.58)	0.46 (0.42, 0.49)	0.56 (0.54, 0.58)
T19	Respiratory Symptoms	T24	Respiratory Symptoms	0.61 (0.59, 0.63)	**0.64** (0.62, 0.66)	0.55 (0.53, 0.57)	0.53 (0.51, 0.55)	0.52 (0.50, 0.54)	0.62 (0.60, 0.64)	0.52 (0.50, 0.53)	0.63 (0.61, 0.65)	0.46 (0.44, 0.48)	0.44 (0.42, 0.46)	0.57 (0.55, 0.59)	0.52 (0.50, 0.54)
T21	COVID-19 Detection	T25	COVID-19 Detection	0.56 (0.54, 0.59)	**0.69** (0.67, 0.71)	0.54 (0.52, 0.56)	0.56 (0.54, 0.58)	0.48 (0.45, 0.50)	0.64 (0.62, 0.66)	0.68 (0.66, 0.70)	0.47 (0.45, 0.49)	0.42 (0.40, 0.44)	0.65 (0.63, 0.67)	0.48 (0.45, 0.50)	0.68 (0.66, 0.70)
T25	COVID-19 Detection	T27	Vocal pathology detection	0.53 (0.50, 0.56)	0.37 (0.34, 0.39)	0.55 (0.52, 0.58)	0.48 (0.45, 0.51)	0.45 (0.41, 0.48)	**0.64** (0.62, 0.67)	0.51 (0.48, 0.52)	0.53 (0.50, 0.55)	0.42 (0.38, 0.45)	0.53 (0.52, 0.55)	0.58 (0.55, 0.61)	0.63 (0.61, 0.66)
T24	Respiratory Symptoms	T27	Vocal pathology detection	0.47 (0.44, 0.50)	0.44 (0.42, 0.47)	0.58 (0.55, 0.61)	0.52 (0.50, 0.55)	0.42 (0.38, 0.46)	0.63 (0.60, 0.65)	0.57 (0.54, 0.59)	0.57 (0.55, 0.59)	0.49 (0.46, 0.52)	0.58 (0.57, 0.60)	**0.64** (0.62, 0.67)	0.63 (0.61, 0.66)
T25	COVID-19 Detection	T21	COVID-19 Detection	0.54 (0.50, 0.58)	0.61 (0.57, 0.65)	0.55 (0.51, 0.58)	0.51 (0.47, 0.54)	0.48 (0.44, 0.52)	0.57 (0.53, 0.61)	0.57 (0.52, 0.61)	**0.61** (0.57, 0.65)	0.51 (0.47, 0.55)	0.57 (0.53, 0.61)	0.55 (0.51, 0.59)	0.54 (0.50, 0.58)
Cross-category															
Conceptualization		Formulation		0.66 (0.60, 0.72)	0.49 (0.43, 0.55)	**0.67** (0.61, 0.72)	0.33 (0.29, 0.39)	0.30 (0.26, 0.36)	0.61 (0.57, 0.65)	0.27 (0.22, 0.32)	0.60 (0.54, 0.66)	0.55 (0.48, 0.60)	0.26 (0.21, 0.30)	0.51 (0.45, 0.57)	0.27 (0.23, 0.30)
Conceptualization		Articulation (Neuromuscular)		0.46 (0.46, 0.47)	0.57 (0.56, 0.57)	0.57 (0.56, 0.57)	**0.71** (0.70, 0.71)	0.64 (0.64, 0.65)	0.61 (0.61, 0.62)	0.51 (0.51, 0.52)	0.58 (0.57, 0.59)	0.61 (0.61, 0.62)	0.30 (0.29, 0.31)	0.51 (0.50, 0.51)	0.64 (0.63, 0.64)
Conceptualization		Articulation (Phonatory / Respiratory)		0.48 (0.46, 0.49)	0.50 (0.49, 0.51)	0.55 (0.54, 0.57)	0.48 (0.46, 0.49)	0.48 (0.47, 0.50)	**0.60** (0.59, 0.62)	0.49 (0.48, 0.51)	0.45 (0.44, 0.46)	0.50 (0.49, 0.52)	0.44 (0.43, 0.46)	0.45 (0.44, 0.46)	0.55 (0.54, 0.56)
Formulation		Conceptualization		0.51 (0.49, 0.54)	0.46 (0.44, 0.49)	0.52 (0.49, 0.55)	0.53 (0.51, 0.56)	0.44 (0.42, 0.46)	**0.74** (0.71, 0.76)	0.48 (0.47, 0.51)	0.53 (0.51, 0.55)	0.52 (0.50, 0.54)	0.49 (0.47, 0.52)	0.47 (0.45, 0.50)	0.45 (0.43, 0.47)
Formulation		Articulation (Neuromuscular)		0.63 (0.62, 0.63)	0.73 (0.72, 0.73)	**0.80** (0.80, 0.81)	0.80 (0.79, 0.80)	0.37 (0.37, 0.38)	0.54 (0.53, 0.54)	0.59 (0.58, 0.60)	0.59 (0.59, 0.60)	0.71 (0.71, 0.72)	0.52 (0.51, 0.52)	0.52 (0.51, 0.53)	0.46 (0.46, 0.47)
Formulation		Articulation (Phonatory / Respiratory)		0.51 (0.50, 0.52)	0.53 (0.52, 0.54)	0.58 (0.57, 0.59)	0.53 (0.52, 0.55)	0.53 (0.51, 0.54)	**0.58** (0.57, 0.59)	0.54 (0.53, 0.56)	0.53 (0.52, 0.54)	0.54 (0.53, 0.55)	0.52 (0.51, 0.54)	0.54 (0.53, 0.55)	0.49 (0.48, 0.50)
Articulation (Neuromuscular)		Conceptualization		0.52 (0.51, 0.53)	0.54 (0.53, 0.55)	0.56 (0.55, 0.58)	**0.57** (0.56, 0.58)	0.52 (0.51, 0.54)	0.49 (0.48, 0.50)	0.51 (0.50, 0.52)	0.55 (0.54, 0.56)	0.57 (0.55, 0.58)	0.43 (0.41, 0.44)	0.51 (0.50, 0.52)	0.56 (0.55, 0.57)
Articulation (Neuromuscular)		Formulation		0.81 (0.79, 0.84)	0.86 (0.84, 0.88)	0.73 (0.70, 0.76)	0.48 (0.45, 0.50)	0.36 (0.33, 0.38)	0.51 (0.48, 0.55)	0.86 (0.85, 0.88)	0.82 (0.80, 0.84)	0.73 (0.70, 0.76)	**0.87** (0.86, 0.88)	0.82 (0.80, 0.84)	0.76 (0.74, 0.79)
Articulation (Neuromuscular)		Articulation (Phonatory / Respiratory)		0.53 (0.52, 0.54)	0.54 (0.53, 0.55)	0.55 (0.54, 0.56)	0.53 (0.53, 0.54)	0.51 (0.50, 0.52)	**0.59** (0.58, 0.59)	0.55 (0.54, 0.56)	0.53 (0.52, 0.54)	0.53 (0.52, 0.53)	0.55 (0.55, 0.56)	0.56 (0.55, 0.57)	0.54 (0.54, 0.55)
Articulation (Phonatory / Respiratory)		Conceptualization		0.46 (0.43, 0.48)	0.45 (0.42, 0.49)	0.50 (0.47, 0.55)	0.48 (0.45, 0.51)	0.42 (0.40, 0.44)	**0.83** (0.82, 0.85)	0.50 (0.47, 0.54)	0.47 (0.45, 0.51)	0.46 (0.44, 0.48)	0.46 (0.44, 0.50)	0.45 (0.42, 0.48)	0.57 (0.54, 0.63)
Articulation (Phonatory / Respiratory)		Formulation		0.64 (0.59, 0.68)	0.79 (0.75, 0.82)	0.73 (0.68, 0.77)	0.80 (0.75, 0.84)	0.67 (0.61, 0.73)	0.61 (0.57, 0.65)	**0.88** (0.85, 0.90)	0.73 (0.69, 0.77)	0.55 (0.49, 0.61)	0.81 (0.77, 0.85)	0.57 (0.51, 0.63)	0.86 (0.83, 0.89)
Articulation (Phonatory / Respiratory)		Articulation (Neuromuscular)		0.63 (0.63, 0.64)	0.59 (0.59, 0.60)	0.67 (0.67, 0.68)	0.72 (0.71, 0.73)	0.52 (0.51, 0.53)	0.68 (0.68, 0.69)	0.75 (0.74, 0.75)	0.55 (0.54, 0.56)	0.53 (0.53, 0.54)	0.65 (0.64, 0.65)	0.59 (0.58, 0.59)	**0.75** (0.74, 0.76)

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state two primary contributions: (1) SpeechDx, a benchmark of 12 datasets and 27 tasks organized by the speech production mechanism, and (2) a systematic evaluation of 12 audio encoders in-domain and under zero-shot cross-condition transfer. Both contributions are realized in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated section (Section 6) discussing the limitations of this work. First, we acknowledge that the objectivity of clinical labels varies across the 27 tasks: many datasets compare healthy controls with patients who have lived with a condition for years rather than at the early-symptom stage relevant for screening, and several severity scores rely on self- or clinician-reported instruments subject to anchoring and recall bias. Second, the benchmark is primarily composed of English-language recordings and contains uneven representation of accents, ages, and gender, which limits generalization claims to broader populations and motivates demographic-stratified analyses as future work. Finally, the benchmark covers 13 datasets selected to span the four stages of speech production, but several clinically relevant conditions (e.g., Huntington's disease, pediatric speech sound disorders) are not represented because suitable public datasets are not yet available; the taxonomy is designed to accommodate such datasets as they become available. We frame these limitations as scoping decisions and concrete directions for future extension of the benchmark. Notably, despite these constraints, current state-of-the-art speech/audio models fail to reliably solve these foundational tasks, underscoring the need to establish robust performance on this benchmark before progressing to more complex real-world screening scenarios.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution

is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: We do not present theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides comprehensive information to enable reproducibility of all experimental results. First, we release the complete benchmark codebase, including dataset processing pipelines, model training scripts, and evaluation protocols at `https://anonymous.4open.science/r/SpeechDx-F584`. Second, Section 4 and Appendix D provides detailed specifications of all model architectures, hyperparameters, training procedures, and data preprocessing steps. Third, for each of the 13 datasets in the benchmark, we document the train/validation/test splits, exclusion criteria, and task formulations in Section 3, Appendix A and Appendix B. For proprietary models (Whisper, HuBERT, etc.), we specify the exact model versions used in Appendix C. The combination of public data, released code, and detailed methodological documentation ensures that all main experimental results can be independently reproduced.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide open access to both code and data with comprehensive reproduction instructions. The complete benchmark codebase is released at `https://anonymous.4open.science/r/SpeechDx-F584`, including: (1) data processing scripts for all 13 datasets, (2) preprocessing pipelines that transform raw audio recordings into train/validation/test splits matching those used in our experiments, (3) model training scripts with exact hyperparameters and environment specifications; (4) evaluation scripts that compute all reported metrics and generate tables/figures from the paper; and (5) a detailed README with instructions to reproduce each experimental result. All 13 datasets are publicly available from their original sources (see Table 2).

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides comprehensive experimental details at multiple levels of granularity. Section 4 specifies the core experimental setup including training procedures, and evaluation protocols. For each dataset, we document the train/validation/test splits in Section 3, with sample counts provided in Table 1. Hyperparameters selection and optimization is described in Appendix D. We further provide a benchmarking codebase that specifies additional information to reproduce our results at `https://anonymous.4open.science/r/SpeechDx-F584`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Given the scale of our benchmark (12 models evaluated across 27 tasks, with additional zero-shot evaluations), training each model multiple times with different random seeds is computationally prohibitive. Instead, we quantify uncertainty through bootstrap resampling of test set predictions, which captures variability due to finite sample size (refer to Section 4. Confidence intervals are reported in all results tables and key figures.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide detailed compute resource information in Section 4.4.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research fully conforms to the NeurIPS Code of Ethics. We have carefully reviewed the guidelines and ensured compliance in all aspects. Our work uses only publicly available datasets with appropriate citations and respects original data usage licenses. All human subjects data comes from previously published datasets with proper IRB approval and informed consent obtained by the original data collectors.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper includes a discussion of broader impacts in Section 6. We discuss several positive impacts: improved accessibility of healthcare screening through low-cost voice-based tools, potential for early detection of respiratory and neurological conditions in underserved populations, and democratization of health monitoring through commodity devices. Regarding negative impacts, we note that while this is a research benchmark and not a system proposed for deployment, future models developed using this benchmark could face risks if deployed prematurely without clinical validation. We identify key concerns for such future work: demographic bias, privacy considerations, and the critical need for regulatory approval before any patient-facing applications. We explicitly caution that benchmark performance does not imply clinical readiness and recommend demographic-stratified evaluation as important future work.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The benchmark does not release new datasets or pre-trained models with high misuse risk. We provide evaluation code and standardized task definitions for 13 existing public datasets, all of which have been previously released by their original authors with appropriate ethical review and data use agreements. We do not scrape new data from the Internet or release any trained models. The benchmark code includes documentation directing users to obtain dataset access through official channels where data use agreements are required.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets, models, and code libraries used in this work are properly credited with full citations to original papers. For datasets requiring data use agreements, we note the access procedure and direct users to the official sources (see Table 2).

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The benchmark codebase includes comprehensive documentation: a README with installation and usage instructions, detailed task specifications in Section 3, processing pipeline documentation for each dataset, and complete evaluation protocol descriptions in Section 4. We do not collect new data; informed consent for all datasets was obtained by original data collectors as documented in their publications.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or new human subjects research. All datasets are existing publicly available collections where human subjects protocols were handled by the original data collectors.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not conduct new research with human subjects. All datasets are publicly available collections where IRB approval and informed consent were obtained by the original data collectors as documented in their respective publications.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: With the exception of the audio encoders that we evaluated on our benchmark, which were written and released by their respective creators, our core experimental codebase was written entirely on our own by hand. LLMs were only used for peripheral affordances such as refactoring and documentation.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.