Something from Nothing: Data Augmentation for Robust Severity Level Estimation of Dysarthric Speech

Jaesung Bae [1,*], *Xiuwen Zheng* [1,*], *Minje Kim* [1], *Chang D. Yoo* [2], *Mark Hasegawa-Johnson* [1,**]

[1] University of Illinois Urbana-Champaign, IL, USA
[2] Korea Advanced Institute of Science & Technology, KR

{jb82, xiuwenz2, minje, jhasegaw}@illinois.edu, cd_yoo@kaist.ac.kr

Abstract

Dysarthric speech quality assessment (DSQA) is critical for clinical diagnostics and inclusive speech technologies. However, subjective evaluation is costly and difficult to scale, and the scarcity of labeled data limits robust objective modeling. To address this, we propose a three-stage framework that leverages unlabeled dysarthric speech and large-scale typical speech datasets to scale training. A teacher model first generates pseudo-labels for unlabeled samples, followed by weakly supervised pretraining using a label-aware contrastive learning strategy that exposes the model to diverse speakers and acoustic conditions. The pretrained model is then fine-tuned for the downstream DSQA task. Experiments on five unseen datasets spanning multiple etiologies and languages demonstrate the robustness of our approach. Our Whisper-based baseline significantly outperforms SOTA DSQA predictors such as SpICE, and the full framework achieves an average SRCC of 0.761 across unseen test datasets.

Index Terms: dysarthria, speech quality assessment, data augmentation, weakly-supervised, contrastive learning

1. Introduction

Dysarthria is a motor speech disorder caused by neurological impairments, leading to substantial degradation in acoustic and perceptual characteristics. Accurate dysarthric speech quality assessment (DSQA) is essential for clinical diagnosis, early detection of progressive neurological conditions, rehabilitation monitoring, and the development of inclusive speech technologies, including pathological speech enhancement and automatic speech recognition. However, DSQA relies on expert clinical evaluation by speech-language pathologists (SLPs), limiting its scalability and accessibility in real-world settings. An automated and objective DSQA system would therefore complement SLP expertise by enabling continuous monitoring outside the clinic.

Recent deep learning-based SQA models can reliably assess the perceptual correlates of algorithmic and channel distortion; in particular, non-intrusive SQA (NI-SQA) approaches estimate speech quality directly from the input signal without reference recordings. For example, DNSMOS [1] evaluates speech quality in real-world noisy and reverberant conditions, while UTMOS [2] predicts subjective quality for clean and enhanced speech. The growing availability of large-scale dysarthric datasets has further enabled the development of deep learning-based DSQA models. SpICE [3] is trained on 550,000 disordered speech samples from *Project Euphonia* [4] to automatically predict human-perceived intelligibility. More recently, [5] develops voice quality probes for dysarthric speech across seven perceptual dimensions, trained on 11,184 samples from the *Speech Accessibility Project (SAP)* [6]. These previous approaches show promising results; however, most evaluations have been conducted on English corpora, which may limit their generalizability to non-English datasets.

Contrastive learning methods such as SimCLR [7] are a powerful self-supervised learning method that can form a structured latent representation space. It has also been widely adopted in speech technology, e.g., to train speech foundation models such as wav2vec 2.0 [8] and HuBERT [9]. This representation space is also proven to be effective for DSQA approaches, such as in [3, 5]. However, to transfer the speech foundation model into the downstream DSQA task, a labeled dataset is required, and the availability of in-domain training data remains scarce. The Euphonia dataset [4] is not publicly released, limiting its use in research. In the SAP corpus [6], only a small fraction of each participant's speech, i.e., roughly 30 read sentences, is rated by speech-language pathologists, compared with 350–400 read sentences and 50–80 spontaneous speech samples in total. This limited amount of labeled datasets for dysarthric speech may limit the performance of DSQA models, especially in terms of robustness.

To fully leverage the SAP dataset, which contains a small labeled subset and a much larger unlabeled corpus, we propose a three-stage framework. Additionally, we propose to adopt the large-scale typical speech corpus LibriSpeech [10] to enhance the speaker variability and acoustic environment of the training dataset and improve the robustness. In the first stage, a teacher regression model is trained on the limited labeled SAP subset using Whisper-large [11] as the speech encoder, then used to generate pseudo-labels for the extensive unlabeled SAP samples. In the second stage, these pseudo-labeled samples are then combined with Librispeech and used for weakly-supervised pretraining. During this stage, we employ a label-aware contrastive learning approach inspired by [12] to better align the representations with perceptual quality labels. In the final stage, the pretrained representation layers are fine-tuned on the labeled SAP subset for the downstream regression task.

To demonstrate the effectiveness of our proposed methods, especially in terms of robustness to unseen test data, we build diverse test datasets with varying etiologies and languages, including UASpeech [13], DysArinVox [14], EasyCall [15], EWA-DB [16], and NeuroVoz [17]. Our baseline model, trained on the SAP labeled subset, achieves an utterance-level Spearman's Rank Correlation Coefficient (SRCC) of 0.719 on the SAP test set and an average speaker-level SRCC of 0.732 on

[*] These authors contributed equally.
[**] indicates the corresponding author.

the cross-domain test sets. Our proposed three-stage framework consistently improves over the baseline, reaching an average SRCC of 0.761 on the cross-domain test sets while preserving performance on the SAP dataset. We further demonstrate that weak supervision is essential for harmonizing the SAP and LibriSpeech datasets and improving cross-domain robustness. Our contributions are summarized as follows.

- We propose a three-stage framework that gradually improves the labeling quality of the SAP dataset by using pseudo-labeling, followed by representation learning with a contrastive objective. This method fully leverages the large portion of unlabeled data in the SAP dataset.
- We propose to incorporate a large-scale typical speech dataset, LibriSpeech, into our DSQA model training pipeline, which improves the speaker and acoustic environment variability.
- We evaluate our method with various cross-domain dysarthric speech datasets with diverse labeling metrics and languages, demonstrating the effectiveness and robustness of our proposed methods.
- Through a rigorous ablation study, we demonstrate the importance of providing weak supervision during the representation learning stage and adopting the LibriSpeech dataset.
- Model checkpoint and training code are publicly available: `https://github.com/JaesungBae/DA-DSQA`

2. Related Works

2.1. Speech Quality Assessment

Speech quality assessment (SQA) aims to predict perceived speech quality, typically represented by a mean opinion score (MOS). Traditional intrusive metrics such as PESQ [18] and signal-based measures like SI-SNR [19] and STOI [20] require reference signals, and even codec-oriented intrusive metrics like WARP-Q [21] share this limitation, whereas modern generative applications (e.g., text-to-speech (TTS) and speech enhancement) increasingly rely on non-intrusive SQA (NI-SQA) methods. With the availability of large-scale MOS datasets and benchmarks, deep learning approaches have become the dominant paradigm for NI-SQA. Representative systems include DNSMOS [1], which predicts multi-dimensional MOS for the deep noise suppression (DNS) task, and UTMOS [2], which employs a strong-weak learner framework to model human perception of the naturalness of synthetic speech and achieved state-of-the-art performance in the VoiceMOS Challenge 2022 [22]. More recently, [23] introduced a training-free NI-SQA metric based on neural codec latent representations. While DNSMOS and UTMOS have been applied to evaluate synthetic dysarthric speech [24, 25], they are primarily designed for healthy or synthetic speech, leaving their applicability to pathological speech unclear.

2.2. Dysarthric Speech Quality Assessment

Unlike conventional SQA, perceptual quality assessment of dysarthric speech often focuses on clinically relevant attributes such as intelligibility and other voice quality dimensions assessed by speech-language pathologists (SLPs). SpICE [3] automatically estimates human-perceived intelligibility of dysarthric speech using large-scale data from Project Euphonia [4]. More recently, [5] proposes voice quality probes trained on Speech Accessibility Project (SAP) data to predict multiple perceptual dimensions of dysarthric speech, including naturalness and intelligibility.

While these studies show that perceptual attributes of pathological speech can be learned from large-scale datasets using self-supervised representations, most approaches rely on foundation models [8, 9, 26] trained predominantly on healthy speech. As a result, the learned representations may be suboptimal for capturing dysarthric speech characteristics, highlighting the need for representations more sensitive to dysarthric speech.

2.3. Contrastive Learning

Contrastive learning is a powerful approach for representation learning that pulls similar samples together while pushing dissimilar ones apart. Self-supervised methods such as SimCLR [7] learn representations from unlabeled data by making augmented pairs of the same sample similar while treating others as negatives. This process enables the model to capture the intrinsic structure of the data. In speech, contrastive objectives underpin self-supervised models such as wav2vec 2.0 [8] and HuBERT [9], which show strong transferability across downstream tasks [27, 28] including DSQA [3, 5]. However, purely self-supervised contrastive learning does not explicitly align representations with task-relevant perceptual attributes. Supervised contrastive learning (SupCon) [12] addresses this limitation by defining samples within the same label as positive pairs. [29] further improves this approach for the regression task by contrasting samples based on the label order, and improves robustness, efficiency, and generalization. [30] proposes an SQA model with contrastive pretraining on audio pairs generated by injecting noise at perceptually similar SNR levels. However, applying such methods to DSQA is challenging due to highly limited and imbalanced labels.

Motivated by these advances, we propose a weakly supervised contrastive learning strategy that leverages pseudo-label-informed similarity together with coarse pairing between typical and dysarthric speech. This enables task-aware representation learning while mitigating label noise from a label-constrained dysarthric speech dataset.

3. Background

3.1. Severity Level Prediction

Severity-level prediction is a key dimension of dysarthric speech quality assessment (DSQA), aiding clinical assessment and providing auxiliary supervision for downstream tasks such as automatic dysarthric speech recognition (ASR) [31] and dysarthric speech generation with a TTS model [32]. However, collecting dysarthric speech data with severity-level annotations is costly as it requires expert clinical evaluation. Moreover, existing datasets suffer from labeling scale discrepancies and class imbalance, thus exhibiting inherently ambiguous decision boundaries between severity levels. To address these challenges, we formulate severity prediction as a regression task rather than classification, enabling soft estimation of speech impairment levels, following prior work [5, 33].

Table 1 shows the summary of the training and test datasets used in this work. For training, we use the SAP dataset [6] (covering data collected and annotated through January 31, 2025). Although SAP is one of the largest available dysarthric speech corpora, human expert-annotated samples, based on clinically or perceptually relevant metrics, remain limited, as noted in Section 1. Out of the dataset's 32 perceptual rating dimensions, we select *naturalness* and *intelligibility* as the primary indicators of severity. Each is rated on a 7-point scale, where 1 denotes

Table 1: *Summary of dataset information. Training is English-only, while cross-domain tests are multi-lingual. EN: English; ZH: Mandarin Chinese; IT: Italian; CS: Czech; SK: Slovak; ES: Spanish. ALS: Amyotrophic lateral sclerosis; CP: Cerebral palsy; DS: Down syndrome; PD: Parkinson's disease; AD: Alzheimer's disease; MCI: Mild Cognitive Impairment. DysArinVox provides fine-grained annotations across 20 disorder categories (e.g., paralysis, leukoplakia, and polyps), and EasyCall includes dysarthria associated with PD, Huntington's disease (HD), ALS, peripheral neuropathy, and myopathic or myasthenic lesions. LBL and Spk indicate the labeling method and the number of speakers, respectively.*

Lang.	Name	Etiology	LBL	Hours	# Spk
Training Data					
EN	SAP [6] Train	ALS, CP, DS, PD	✓	10.8	318
			✗	232.9	722
EN	LibriSpeech [10]	Healthy	–	921.7	2.3k
In-domain Evaluation Data					
EN	SAP [6] Test	ALS, CP, DS, PD	✓	3.1	89
Cross-domain Evaluation Data					
EN	UASpeech [13]	CP	Intellig.	7.8	15
ZH	DysArinVox [14]	Mixed	MOS	2.3	72
IT	EasyCall [15]	Mixed	TOM	10.0	44
CS/SK	EWA-DB [16]	PD, AD, MCI	MoCA	4.4	136
ES	NeuroVoz [17]	PD	H-Y	1.7	98

typical speech, and 7 indicates severe dysarthria. We define the target severity score as the average of these two ratings. Considering only utterances shorter than 15 seconds, 10.8 hours of speech are labeled, which is only about 4.6% of the 232.9 hours of unlabeled speech. The training splits of the LibriSpeech [10] dataset are used as an additional normal speech dataset for contrastive learning.

In addition to the SAP in-domain test, we further evaluate generalizability using cross-domain datasets: UASpeech [13], DysArinVox [14], EasyCall [15], EWA-DB [16], and NeuroVoz [17]. These dysarthric speech corpora span different languages and etiologies, and adopt labeling schemes that differ substantially from SAP, making cross-dataset severity prediction extremely challenging. Specifically, UASpeech provides overall speech intelligibility scores derived from word-recognition accuracy by five native listeners; DysArinVox reports MOS ratings based on subjective perceptual evaluations of Mandarin dysphonic and dysarthric speech; EasyCall provides therapy outcome measure (TOM) scores assigned by experienced neurologists to categorize the clinical severity levels; EWA-DB provides Montreal cognitive assessment (MoCA) scores reflecting cognitive health status; NeuroVoz provides Hoehn and Yahr (H-Y) scale stages to categorize the progressive severity of Parkinson's disease symptoms and their impact on motor function.

As these cross-domain datasets are labeled per speaker rather than per utterance, we predict utterance-level severity scores and average them per speaker. For EasyCall, we merge the official train and test splits to create a larger evaluation set with more speakers. For EWA-DB, we restrict evaluation to speakers with AD, PD, or MCI. Hyperparameters are selected solely based on the validation splits of SAP and EasyCall.

3.2. Contrastive Learning

Our approach builds upon the SimCLR [7] contrastive learning frameworks, and the extended version of it with label supervision, SupCon [12]. Both methods generate two augmented views of each sample within a mini-batch, resulting in two copies of the batch. In our framework, to reduce the computational burden of performing inference on the Whisper-large model at each iteration, the augmentation process is defined in the feature vectors extracted by the Whisper encoder. Let the extracted feature vector be $H_i = \text{Whisper}(\mathbf{x}_i)$. For a randomly sampled mini-batch of size B, we obtain $\{H_i, y_i\}_{i=1}^{B}$, where H_i is normalized. Note that y_i may correspond to the pseudo-label $\bar{y}_i$ for pseudo-labeled data; for simplicity, we denote both as y_i. We construct an augmented batch $\{\tilde{H}_j, \tilde{y}_j\}_{j=1}^{2B}$, where $\tilde{H}_i$ and $\tilde{H}_{B+i}$ are two independently augmented versions of H_i for $i = 1, \ldots, B$, and $\tilde{y}_i = \tilde{y}_{B+i} = y_i$. The final representation $\mathbf{z}_i$ is then obtained by passing $\tilde{H}_i$ through two linear layers, followed by statistical temporal pooling and a final linear layer (Figure 1).

SimCLR [7] is a self-supervised method that treats two different augmented versions of the same representation as a positive pair, while all other samples in the batch serve as negative pairs. Let $i \in I \coloneqq \{1, \ldots, 2B\}$ denote the index of a sample in the augmented batch. For a given index i, there is always a positive pair for it within the same batch, whose index $k(i)$ is defined as the index of the other augmented sample originating from the same source sample. The normalized temperature-scaled cross-entropy (NT-Xent) loss is defined as

$$\mathcal{L}_{\text{SimCLR}} = -\frac{1}{2B} \sum_{i \in I} \log \frac{\exp\left(\mathbf{z}_i^\top \mathbf{z}_{k(i)}/\tau\right)}{\sum_{j \in A(i)} \exp\left(\mathbf{z}_i^\top \mathbf{z}_j/\tau\right)} \tag{1}$$

where $\tau > 0$ is a temperature parameter and $A(i) \coloneqq I \setminus \{i\}$.

However, $\mathcal{L}_{\text{SimCLR}}$ does not utilize label information y_i, although it could be informative for the downstream task. Instead, relies solely on the inherent data structure and cannot explicitly guide the representation space toward task-relevant features. To this end, SupCon proposes a method to consider the data samples with the same label as positive pairs, rather than considering all the other data samples, but the augmented version of it as negative. The positive pair set for index i is defined as follows:

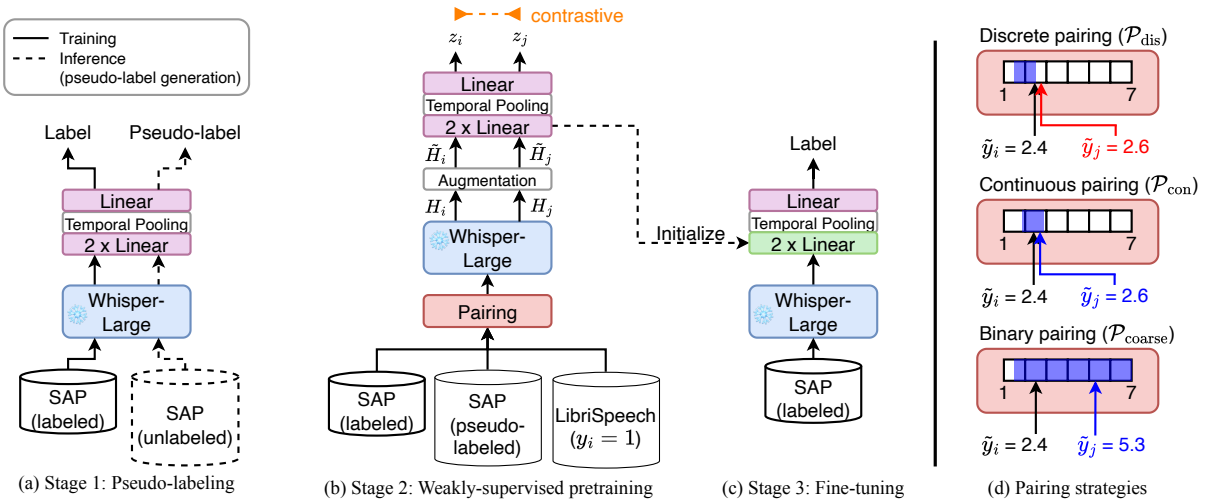
$$\mathcal{P}_{\text{sup}}(i) = \{j \in A(i) \mid \tilde{y}_j = \tilde{y}_i\}\,. \tag{2}$$

Using these positive pairs, SupCon optimizes the model with the following modified objective based on $\mathcal{L}_{\text{SimCLR}}$:

$$\mathcal{L}_{\text{sup}} = -\frac{1}{2B} \sum_{i \in I} \frac{1}{|\mathcal{P}_{\text{sup}}(i)|} \sum_{p \in \mathcal{P}_{\text{sup}}(i)} \log \frac{\exp\left(\mathbf{z}_i^\top \mathbf{z}_p/\tau\right)}{\sum_{j \in A(i)} \exp\left(\mathbf{z}_i^\top \mathbf{z}_j/\tau\right)}. \tag{3}$$

4. Proposed Method

Our goal is to develop a robust and generalizable DSQA model. Due to the scarcity of labeled dysarthric speech data, our key motivation is to leverage large amounts of unlabeled dysarthric speech alongside large-scale typical speech. This allows the model to be exposed to diverse speaker identities and acoustic environments. However, effectively utilizing such unlabeled data is challenging because the absence of reliable severity annotations prevents direct supervised optimization. We propose a three-stage framework: (1) pseudo-label generation using a supervised regression model trained on labeled data, (2) weakly supervised representation learning via contrastive objectives, and (3) fine-tuning a regression model on labeled data.



(a) Stage 1: Pseudo-labeling (b) Stage 2: Weakly-supervised pretraining (c) Stage 3: Fine-tuning (d) Pairing strategies

Figure 1: *(a–c) Illustration of the three-stage framework with weakly supervised pretraining, and (d) the proposed pairing strategies for weakly supervised contrastive learning. (a)* ***Stage 1:*** *A regression model is trained on the labeled SAP dataset (3% of the total), and pseudo-labels are generated for the unlabeled portion. (b)* ***Stage 2:*** *Three linear layers are trained with weakly supervised contrastive losses. Pseudo-labeled SAP data from Stage 1 and LibriSpeech are additionally used, with LibriSpeech assigned a label of one (healthy). (c)* ***Stage 3:*** *The first two linear layers from Stage 2 are fine-tuned with an additional final linear layer using the labeled SAP dataset. (d) In* ***Stage 2****, positive pairs for contrastive learning are generated in three ways: discretizing the label, thresholding the distance between continuous labels, or dividing the label range into a binary dichotomy. With anchor $\tilde{y}_i = 2.4$, samples with $\tilde{y}_j$ in the blue region form positive pairs: for example, $\tilde{y}_j = 2.6$ is negative under discrete pairing but positive under continuous pairing. Binary pairing divides the data into two groups, providing weaker supervision.*

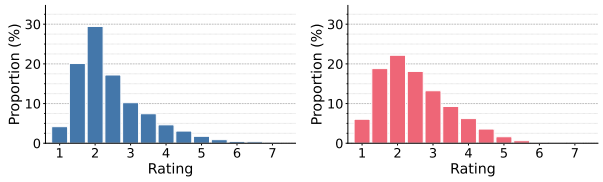


Figure 2: *Histogram of label proportions in $\mathcal{D}_{labeled}$ (left) and $\mathcal{D}_{pseudo}$ (right).*

4.1. Stage 1: Pseudo-Label Generation

We first train a regression model using the labeled subset of SAP. Let $\mathbf{x}$ denote an input speech utterance and $y \in [1, 7]$ its corresponding severity label. Denote the labeled and unlabeled datasets as

$$\mathcal{D}_{\text{labeled}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}, \quad \mathcal{D}_{\text{unlabeled}} = \{\mathbf{x}_i\}_{i=1}^{M},$$

where $M \gg N$. Our regression model architecture is shown in Figure 1. We use the encoder portion of Whisper-large as a speech foundation model to extract feature representations from the input waveform, which is frozen during training. The extracted frame-level features are passed to the trainable component: two linear layers followed by statistical temporal average pooling [34]. The pooled representation is then mapped to a scalar severity score via a final linear layer. After training the regression model on $\mathcal{D}_{\text{labeled}}$, we use it to predict severity scores for $\mathcal{D}_{\text{unlabeled}}$, producing pseudo-labels $\bar{y}_i$. This yields a pseudo-labeled dataset:

$$\mathcal{D}_{\text{pseudo}} = \{(\mathbf{x}_i, \bar{y}_i)\}_{i=1}^{M}.$$

The proportion histogram of $\mathcal{D}_{\text{labeled}}$ and the resulting $\mathcal{D}_{\text{pseudo}}$ over y is shown in Figure 2. The initial regression model trained on $\mathcal{D}_{\text{labeled}}$ achieves an SRCC of 0.719 on the SAP test set (Table 2), indicating reasonably reliable predictions. However, these pseudo-labels are inherently imperfect. Training another regression model on $\mathcal{D}_{\text{pseudo}}$ is ineffective because it would largely replicate the data distribution of $\mathcal{D}_{\text{labeled}}$, while inheriting the bias from the initial model instead of fully leveraging the diversity in $\mathcal{D}_{\text{unlabeled}}$. Hence, we propose to use pseudo-labels only to construct weak supervision signals for representation learning in Stage 2.

4.2. Stage 2: Weakly Supervised Representation Learning

The main objective of this stage is to learn a more structured representation space that is both task-relevant and generalizable for robust assessment of dysarthric speech quality. Contrastive learning has been widely adopted for representation learning [7, 12, 29] and has been shown to improve downstream generalization. Inspired by this, we employ contrastive learning to adapt the Whisper encoder's representation space to our target task using $\mathcal{D}_{\text{labeled}}$ and $\mathcal{D}_{\text{pseudo}}$. Although SAP is relatively large, it consists solely of dysarthric speech, which may limit variability in speaker identities and acoustic environments. To further increase diversity, we additionally incorporate LibriSpeech [10], as a large-scale typical speech dataset. Since a label of 1 in SAP indicates speech with no noticeable impairment, we assign the typical speech samples this label:

$$\mathcal{D}_{\text{Libri}} = \{(\mathbf{x}_i, 1)\}_{i=1}^{K}$$

where K denotes the number of typical samples, we analyze the impact of incorporating LibriSpeech in Section 5.3.2.

4.2.1. Proposed weakly supervised contrastive learning

In our pilot study, we observe that the simply applying SimCLR with our mixed datasets $\mathcal{D}_{\text{labeled}}$, $\mathcal{D}_{\text{pseudo}}$, and $\mathcal{D}_{\text{Libri}}$ was unable

to form a meaningful representation space for the downstream task. This is likely because of the different data characteristics between the SAP and LibriSpeech datasets. Applying the SupCon directly is also problematic because of the continuous nature of the predicted score of the regression model for $\mathcal{D}_{\text{pseudo}}$.

To this end, we first propose a supervised contrastive learning method to discretize the label y_i to the nearest integer and define positive pairs among samples sharing the same discretized label:

$$\lfloor y \rfloor = \max\{n \in \mathbb{Z} \mid n \leq y\}. \tag{4}$$

$$\mathcal{P}_{\text{dis}}(i) = \left\{ j \in A(i) \mid \left\lfloor y_j + \tfrac{1}{2} \right\rfloor = \left\lfloor y_i + \tfrac{1}{2} \right\rfloor \right\}. \tag{5}$$

However, this discretization introduces discontinuities near label boundaries. For example, labels 2.4 and 2.6 are expected to be more similar than 2.4 and 1.7, yet the former pair is not considered positive while the latter is. To address this, we propose an alternative pairing strategy based on label distance. Specifically, we define positive pairs as those whose label difference is smaller than a threshold α:

$$\mathcal{P}_{\text{con}}(i) = \{ j \in A(i) \mid |y_j - y_i| < \alpha \}. \tag{6}$$

However, pseudo-labels may be inaccurate, and excessive reliance on them can introduce misleading supervision. Additionally, as shown in Figure 2, the distribution of labels y and $\bar{y}$ are skewed toward lower rating values, meaning that constructing positive pairs from $\mathcal{P}_{\text{dis}}$ and $\mathcal{P}_{\text{con}}$ risks overfitting to low-score samples. To mitigate these issues, we introduce a coarser supervisory signal by grouping samples into two categories: dysarthric speech and typical speech. Given a threshold β, a sample is considered dysarthric if $y_i > \beta$, and typical otherwise. Based on this binary grouping, the positive pair set is defined as

$$\mathcal{P}_{\text{coarse}}(i) = \{ j \in A(i) \mid \mathbf{1}\{y_j > \beta\} = \mathbf{1}\{y_i > \beta\} \}, \tag{7}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function, which equals 1 if the condition inside the braces is true and 0 otherwise. By proposing this coarse binary grouping strategy, we can expect the low severity level samples to act as a bridge between the LibriSpeech and SAP dataset, while high severity level samples will also learn the structured representation space, including the task-relevant information.

For each positive pair, we compute the corresponding loss objectives $\mathcal{L}_{\text{dis}}$, $\mathcal{L}_{\text{con}}$, and $\mathcal{L}_{\text{coarse}}$ by replacing $\mathcal{P}_{\text{sup}}$ with $\mathcal{P}_{\text{dis}}$, $\mathcal{P}_{\text{con}}$, and $\mathcal{P}_{\text{coarse}}$, respectively. The thresholds α and β are set to 0.5 and 1.5.

Temperature value τ: The temperature value τ is an important factor that controls the strength of the supervision of the contrastive learning as suggested in the literature [12, 35]. If τ is small, the loss function encourages hard decisions even on confusing examples, while big τ values allow for such confusions, leading to a more gradually distributed feature space between different severity levels. As we incorporate a new dataset $\mathcal{D}_{\text{Libri}}$, a small τ can lead to a trivial separation between the features of $\mathcal{D}_{\text{Libri}}$ and those of the SAP dataset $\mathcal{D}_{\text{labeled}}$ and $\mathcal{D}_{\text{unlabeled}}$. We investigate different choices of τ and observe that larger values tend to provide more robust feature spaces (Section 5.3.3).

4.2.2. Variance Regularization

In addition to the contrastive objective, we incorporate the variance regularization term from VICReg [36] to prevent representational collapse. This regularizer penalizes embedding dimensions whose batch-wise standard deviation falls below a predefined threshold γ:

$$\mathcal{L}_{\text{var}} = \frac{1}{d} \sum_{k=1}^{d} \max\left(0, \gamma - \sqrt{\text{Var}(\mathbf{z}_{:,k}) + \varepsilon}\right). \tag{8}$$

where $\mathbf{z} \in \mathbb{R}^{2B \times d}$ denotes the embedding matrix for a batch of $2B$ samples with embedding dimension d, and $\mathbf{z}_{:,k}$ represents the k-th feature dimension across the batch. This hinge-style penalty encourages informative and well-dispersed representations across feature dimensions. We set $\gamma = 1.0$.

4.2.3. Final Objective

The overall representation learning objective in Stage 2 is defined as

$$\mathcal{L}_{\text{Stage2}} = \mathcal{L}_{\text{contrast}} + \lambda \mathcal{L}_{\text{var}}, \tag{9}$$

where λ controls the strength of the variance regularization, and $\mathcal{L}_{\text{contrast}}$ can be one of $\mathcal{L}_{\text{dis}}$, $\mathcal{L}_{\text{con}}$, or $\mathcal{L}_{\text{coarse}}$. We set $\lambda = 0.1$.

4.3. Stage 3: Fine-Tuning for Regression

In the final stage, we initialize the first two adaptor layers using the pretrained weights obtained from Stage 2. A randomly initialized linear layer is added on top to predict continuous severity scores. The entire model is then fine-tuned on $\mathcal{D}_{\text{labeled}}$. By decoupling representation learning from regression optimization, our framework leverages large-scale unlabeled data while mitigating the adverse effects of pseudo-label noise.

5. Experiments

5.1. Experimental setup

Whisper-large-v3 [26] is adopted as the backbone for feature extraction and is frozen throughout training. Whisper features are extracted after applying voice activity detection (VAD) [37] to the original speech signals following [38].

For Stage 1, two linear layers are applied, followed by statistical temporal pooling [34], which aggregates frame-level representations into an utterance-level representation. A final linear layer maps the representation to a severity score. ReLU activation and dropout ($p = 0.1$) are applied after each linear layer. The first two layers have output dimensions of 320. Training uses Huber loss ($\delta = 0.5$) with a batch size of 32 and a learning rate of 10^{-4}, optimized with AdamW [39]. To mitigate class imbalance, we apply label-weighted random sampling. The model is trained for 10 epochs, selecting the best checkpoint based on SAP validation SRCC.

For stage 2, to generate an augmented view, two augmented views are generated by composing Gaussian noise ($\sigma = 0.01$) with stochastic time masking (up to 20% of frames) and random temporal cropping ($\geq$70% of the sequence), each applied with 50% probability to the Whisper feature. Two linear layers with dimension 320, temporal pooling, and one linear layer with 128 output dimension are applied to convert $\hat{H}$ into $\mathbf{z}$. Training uses Adam with learning rate 10^{-3} and weight decay 10^{-5}, and we train for 2 epochs. We observe that as training progresses, performance generally drops. We assume that heavy fine-tuning of the Whisper features—which are already rich and well-structured—may harm their effectiveness. For the contrastive learning methods, we conduct experiments with $\tau \in \{0.1, 1.0, 10.0, 50.0, 100.0\}$ and select the optimal value

based on the average SRCC score on the validation sets after Stage 3. For proposed models using $\mathcal{L}_{\text{dis}}$, $\mathcal{L}_{\text{con}}$, and $\mathcal{L}_{\text{coarse}}$, the selected τ values were 1.0, 0.1, 0.1, and 10.0, respectively.

All Stage 3 settings are identical to those in Stage 1. Experiments are repeated with five random seeds for stability.

5.1.1. Evaluation metrics

To evaluate the monotonic and linear consistency of our regression models, we adopt two correlation-based metrics: SRCC and Pearson Correlation Coefficient (PCC). SRCC assesses whether the predicted scores preserve the relative ordering of samples, while PCC quantifies the strength of their linear relationship. Unlike absolute error metrics (e.g., MAE), both SRCC and PCC are invariant to affine transformations (scaling and shifting), making them particularly suitable for cross-domain test cases where label ranges may differ.

5.1.2. Comparison models

`DNSMOS` [1] employs a CNN-based multi-stage self-teaching framework to predict non-intrusive MOS. Three SSL-based comparison models are also considered. `UTMOS` [2] fine-tunes a wav2vec 2.0 [8] encoder with a BLSTM and linear layers for frame-level prediction, with utterance-level scores obtained by averaging. The linear classifier in `SpICE` [3] is built on 12th-layer (768-d) representations from a wav2vec 2.0 backbone, with final scores computed as a label-weighted average of predicted class probabilities [5]. The `HuBERT Probe` [5] is reproduced by training a LASSO regression probe on a HuBERT-large (1024-d) backbone using $\mathcal{D}_{\text{labeled}}$, with VAD-based silence trimming and the same class-weighted sampler as our method. The LASSO ℓ_1 coefficient is set to $1\text{e}{-7}$, selected by in-domain validation over $\{0, 1\text{e}{-7}, 1\text{e}{-6}, 1\text{e}{-5}\}$.

Within our three-stage framework, we define three additional comparison models: `Baseline`, `SimCLR`, and Rank-N-Contrast (`RNC`) [29]. `Baseline` is a regression model fine-tuned on Whisper-large encoder features without pseudo-labeling or contrastive learning. `SimCLR` applies contrastive learning with $\mathcal{L}_{\text{SimCLR}}$ and does not require pseudo-labels. `RNC` uses a regression-oriented contrastive objective that contrasts samples based on their ranking in the label space to learn continuous, order-preserving representations. For `SimCLR` and `RNC`, the temperature τ is selected on the validation set and set to 0.1 and 100.0, respectively.

5.2. Results

The test results on the in-domain SAP dataset and various multilingual cross-domain datasets are shown in Table 2. Overall, our `Baseline` model consistently outperforms existing NI-SQA (`DNSMOS`, `UTMOS`) and DSQA (`SpICE`, `HuBERT Probe`) models on all datasets except NeuroVoz, demonstrating the effectiveness of adapting the estimator to dysarthric severity levels and the robustness of the Whisper-large encoder. Although `DNSMOS` and `UTMOS` perform well in conventional SQA [24, 25], they generalize poorly to dysarthric speech severity prediction. In particular, `DNSMOS` exhibits low correlation with human-rated severity levels (avg SRCC$<$0.2). `UTMOS` shows relatively high correlation on MOS-style labeled datasets (e.g., DysArinVox) but drops significantly on others. Among existing DSQA metrics, `SpICE` is originally designed as a classification model, and converting it into a regression model may introduce result leakage. This highlights the limitations and challenges of adapting it to various datasets. `HuBERT Probe` achieves the best performance on NeuroVoz, possibly due to overfitting to PD speech, but shows a clear gap compared to `Baseline` on most other datasets.

Compared to the `Baseline`, `SimCLR` performs worse on the in-domain test but generally performs better on cross-domain tests. This is expected: since the `Baseline` is only exposed to the SAP dataset during training, it may be overfitted to that dataset. While it achieves reasonable performance on the cross-domain datasets, likely due to the generally high quality of the SAP data, its performance remains suboptimal. By exposing the model to unlabeled SAP and LibriSpeech datasets in Stage 2, we introduce more than 11.5K speakers and 1.1K hours of additional data, thereby enhancing the robustness of the final regression model. The `RNC` model performs best on the in-domain test but worse than the `SimCLR` model on the cross-domain tests. This is because it relies heavily on the ordering of pseudo-labels, which can be very noisy.

With our weakly supervised contrastive learning method (`Proposed` $\mathcal{L}_{\text{coarse}}$), we further improve performance over `SimCLR`. Our method generally enhances results on the cross-domain sets while maintaining comparable performance on the in-domain test; it achieves the best or the second-best scores for all cross-domain tests except EasyCall. For the EasyCall dataset, it performs slightly worse than `SimCLR`, but still improves SRCC by 0.23 compared to the `Baseline` model. Interestingly, with our proposed fine-grained supervised representation learning ($\mathcal{L}_{\text{dist}}$ and $\mathcal{L}_{\text{con}}$), performance on the in-domain test slightly improves; however, cross-domain performance generally degrades compared to the `Baseline`. We suggest that providing strong and fine-grained guidance via pseudo-labeling may harm generalizability by increasing overfitting to the in-domain set. However, when we increase τ for both methods (Figure 4), the average improvement on the cross-domain tests generally increases, highlighting the effectiveness of weak supervision.

The t-SNE [40] plots of the embedding space after the temporal pooling layers are shown in Figure 3. Even without contrastive learning, the feature space exhibits a certain structure due to the powerful Whisper-large speech foundation model (Figure 3a). However, embeddings from the SAP dataset form subclusters and occupy a separate region from the LibriSpeech embeddings, likely due to other task-irrelevant speech attributes such as speaker identity, limiting the robustness of the model. Without supervision (Figure 3b), the embedding space becomes smoother and more harmonized with LibriSpeech, but still does not form a structure aligned with severity levels. In contrast, Figure 3c, 3d, 3e, and 3f show a smooth embedding space where SAP features are ordered based more on the target labels, resulting in a more informative representation for our downstream task. More importantly, Figure 3f presents the smoothest transition from the LibriSpeech portion (blue crosses) to the level 1 examples in SAP, which explains the overall superior performance of `Proposed` $\mathcal{L}_{\text{coarse}}$ to others. This demonstrates the effectiveness of our representation learning. Meanwhile, we note that, with `RNC` and fine-grained supervision, the embedding spaces of the SAP and LibriSpeech datasets are separate, limiting the models' generalizability.

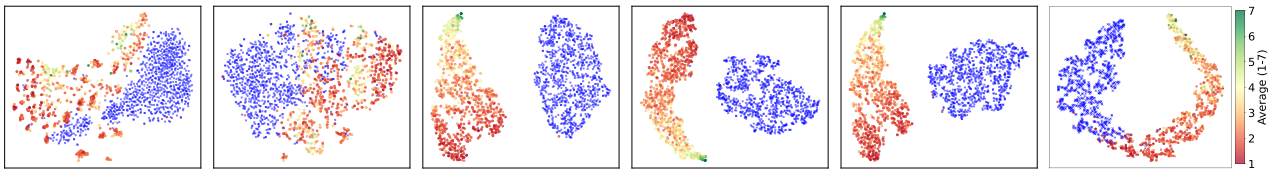
5.3. Ablation Studies

5.3.1. Effectiveness of each stage

Table 3 shows the ablation results for each stage of `Proposed` $\mathcal{L}_{\text{coarse}}$. The first and second rows are nearly identical; without the contrastive learning stage (Stage 2), the model cannot bene-

Table 2: *Performance comparison on in-domain and cross-domain test sets. "Average" in cross-domain indicates the average performance across all cross-domain datasets. Bold and underlined values denote the best and second-best results. Across five seeds, the standard deviations of the in-domain and average cross-domain SRCCs for Proposed ($\mathcal{L}_{coarse}$) are 0.0021 and 0.0041, respectively.*

Model	Stage			In-domain		Cross-domain									
				SAP		UASpeech		DysArinVox		EasyCall		EWA-DB		NeuroVoz		Average	
	1	2	3	SRCC	PCC	SRCC	PCC	SRCC	PCC	SRCC	PCC	SRCC	PCC	SRCC	PCC	SRCC	PCC
DNSMOS [1]				0.186	0.278	0.750	0.742	0.370	0.502	0.041	0.061	-0.274	-0.294	-0.361	-0.363	0.105	0.130
UTMOS [2]				0.489	0.499	0.962	0.912	0.521	0.525	-0.051	-0.108	0.328	0.345	-0.028	-0.065	0.346	0.322
SpICE [3]				0.473	0.505	0.936	0.926	0.538	0.463	0.205	0.237	0.393	0.366	0.180	0.201	0.450	0.439
HuBERT Probe [5]				0.531	0.620	0.927	0.924	0.279	0.334	0.604	0.704	0.588	0.487	**0.705**	**0.706**	0.621	0.631
Baseline			✓	0.719	0.755	0.949	0.963	0.578	0.619	0.849	0.783	0.709	0.686	0.575	0.601	0.732	0.730
SimCLR		✓	✓	0.716	0.739	0.951	0.960	0.628	**0.635**	**0.902**	**0.852**	0.663	0.654	0.576	0.595	0.744	0.739
Rank-N-Contrast	✓	✓	✓	**0.726**	**0.758**	0.959	0.972	0.564	0.601	0.868	0.803	**0.714**	0.693	0.577	0.592	0.736	0.732
Proposed ($\mathcal{L}_{dis}$)	✓	✓	✓	0.724	0.754	0.948	0.962	0.583	0.626	0.838	0.78	0.698	0.673	0.572	0.602	0.728	0.729
Proposed ($\mathcal{L}_{cont}$)	✓	✓	✓	0.722	0.755	0.935	0.956	0.558	0.599	0.874	0.811	0.677	0.659	0.516	0.553	0.712	0.716
Proposed ($\mathcal{L}_{coarse}$)	✓	✓	✓	0.716	0.750	**0.975**	**0.976**	**0.631**	0.632	0.872	0.810	0.711	**0.695**	0.617	0.630	**0.761**	**0.749**



(a) *No contrast* (b) *SimCLR* (c) *Rank-N-Contrast* (d) *Proposed ($\mathcal{L}_{dis}$)* (e) *Proposed ($\mathcal{L}_{con}$)* (f) *Proposed ($\mathcal{L}_{coarse}$)*

Figure 3: *t-SNE figures after stage 2 with various contrastive loss choices. We randomly select 1000 samples in LibriSpeech and SAP training data. These are the embeddings right after the temporal pooling. Blue cross represents the LibriSpeech data, and circles indicate the SAP dataset. From red to green, the color indicates the low-severity to high-severity levels. Best viewed in color.*

fit from the additional pseudo-labeled dataset, as the pseudo-labels essentially imitate the distribution of the SAP-labeled dataset. Conversely, without the pseudo-labeling stage (third row), we have no way but to label the LibriSpeech portion as non-dysarthric, while all the unlabeled SAP data are considered dysarthric, instead of using a threshold β as in eq. (7). This prevents the model from distinguishing speech with low severity levels (i.e., speech close to typical) within the SAP unlabeled dataset, resulting in incorrect supervision of the embedding space and causing misalignment between the SAP and LibriSpeech datasets.

5.3.2. *Effectiveness of $\mathcal{D}_{Libri}$, $\mathcal{D}_{unlabeled}$, and $\mathcal{L}_{var}$*

In Table 4, we observe that $\mathcal{D}_{\text{Libri}}$ plays a crucial role in enhancing model performance. Although performance increases on some cross-domain tests, the average SRCC score remains almost the same as that of the `Baseline` model. This demonstrates that incorporating additional typical speech with diverse speakers and acoustic conditions is effective for improving the robustness of dysarthric speech severity prediction. Excluding $\mathcal{D}_{\text{unlabeled}}$ also degrades performance, but less so than excluding $\mathcal{D}_{\text{Libri}}$, because $\mathcal{D}_{\text{unlabeled}}$ and $\mathcal{D}_{\text{labeled}}$ share overlapping speakers, meaning the model retains partial exposure to those speakers through $\mathcal{D}_{\text{labeled}}$ alone. Without $\mathcal{L}_{\text{var}}$, the overall performance degrades, highlighting its effectiveness in improving robustness in representation learning.

5.3.3. *Temperature value τ*

The SRCC and PCC improvements over the `Baseline` model are shown in Figure 4. For the `Proposed` models with different contrastive losses, performance generally improves as the temperature parameter τ increases. This highlights the importance of weak supervision in contrastive learning for harmonizing the three training datasets: $\mathcal{D}_{\text{labeled}}$, $\mathcal{D}_{\text{pseudo}}$, and $\mathcal{D}_{\text{Libri}}$. While `SimCLR` achieves the best cross-domain performance at $\tau = 10$, its in-domain performance drops significantly, indicating reduced stability compared to the proposed methods.

When τ is small, the contrastive loss emphasizes hard positive and negative pairs. Since SAP and LibriSpeech may differ significantly in their acoustic characteristics, they are easily distinguishable. As a result, a small τ encourages the model to focus on dataset-specific structures rather than learning shared representations across datasets. This effect is also reflected in Figure 5: with a small τ, SAP and LibriSpeech representations are clearly separated in the t-SNE visualization, whereas larger τ values produce more aligned and harmonized embeddings.

These observations suggest that using a larger τ facilitates better integration of the external typical dataset with dysarthric speech data, ultimately improving the robustness of the model.

6. Broad Impact

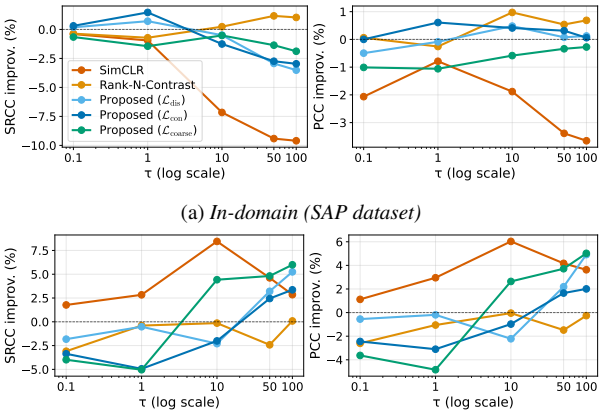
This work advances scalable and automated assessment of dysarthric speech severity, with potential benefits for clinical monitoring, rehabilitation, and the development of more inclusive speech technologies. Our experiments further suggest that the proposed approach can generalize to non-English languages, where labeled data are often even more scarce than in English. Moreover, while our study focuses on dysarthric speech assessment, the proposed framework for leveraging weak supervision and pseudo-labeled data may also be applicable to other domains where labeled data are limited, enabling scalable data augmentation and representation learning beyond speech tasks. While we believe this tool has the potential to positively impact accessibility for individuals with dysarthria, several considerations remain important. Automated predictions

Table 3: *Ablation study of each stage of the Proposed ($\mathcal{L}_{coarse}$) model.*

Index	Stage			In-domain		Cross-domain											
				SAP		UASpeech		DysArinVox		EasyCall		EWA-DB		NeuroVoz		Average	
	1	2	3	SRCC	PCC	SRCC	PCC	SRCC	PCC	SRCC	PCC	SRCC	PCC	SRCC	PCC	SRCC	PCC
1			✓	**0.719**	**0.755**	0.949	0.963	0.578	0.619	0.849	0.783	0.709	0.686	0.575	0.601	0.732	0.730
2	✓		✓	**0.719**	**0.755**	0.949	0.963	0.577	0.620	0.856	0.785	0.709	0.686	0.571	0.598	0.732	0.730
3		✓	✓	0.681	0.740	0.960	0.968	0.604	**0.654**	**0.902**	**0.850**	0.670	0.667	0.596	0.621	0.746	**0.752**
4	✓	✓	✓	0.716	0.750	**0.975**	**0.976**	**0.631**	0.632	0.872	0.810	**0.711**	**0.695**	**0.617**	**0.630**	**0.761**	0.749

Table 4: *Ablation study on additional LibriSpeech dataset, unlabeled portion of SAP dataset, and variance regularization objective.*

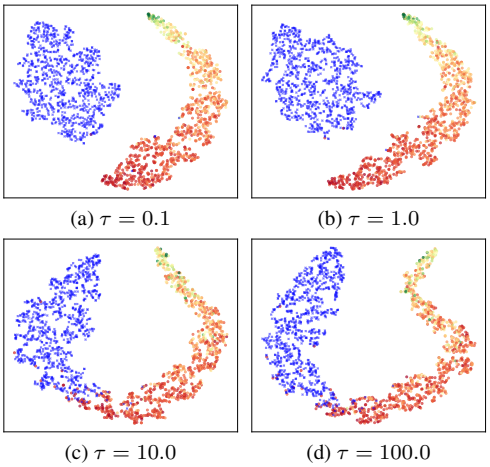
Model	In-domain		Cross-domain											
	SAP		UASpeech		DysArinVox		EasyCall		EWA-DB		NeuroVoz		Average	
	SRCC	PCC	SRCC	PCC	SRCC	PCC	SRCC	PCC	SRCC	PCC	SRCC	PCC	SRCC	PCC
Proposed ($\mathcal{L}_{\text{coarse}}$)	**0.716**	0.750	**0.975**	0.976	0.631	0.632	0.872	0.810	**0.711**	**0.695**	0.617	0.630	**0.761**	**0.749**
w/o $\mathcal{D}_{\text{Libri}}$	0.706	0.743	0.952	0.964	0.605	0.630	0.832	0.746	0.692	0.668	0.590	0.613	0.734	0.724
w/o $\mathcal{D}_{\text{unlabeled}}$	0.715	**0.755**	0.950	0.956	0.609	0.625	**0.877**	0.799	0.681	0.660	**0.632**	**0.659**	0.750	0.740
w/o $\mathcal{L}_{\text{var}}$	0.711	0.753	0.963	**0.978**	**0.661**	**0.658**	0.862	**0.813**	0.683	0.660	0.606	0.616	0.755	0.745



(a) *In-domain (SAP dataset)*

(b) *Cross-domain (average)*

Figure 4: *The improvement percentages of SRCC and PCC over the* `Baseline` *model vary with different values of τ. (a) In-domain testset (SAP dataset) and (b) average scores of cross-domain testsets. In general, the performance of our proposed methods improves as τ increases. Although* `SimCLR` *achieves the best cross-domain average performance at $\tau = 10$, its in-domain test performance deteriorates significantly, highlighting the robustness of our proposed methods.*



(a) $\tau = 0.1$ (b) $\tau = 1.0$

(c) $\tau = 10.0$ (d) $\tau = 100.0$

Figure 5: *Embedding spaces with different τ. Since the LibriSpeech and SAP datasets have distinct characteristics, they can be considered easy positive/negative pairs. Therefore, with a small τ, contrastive learning tends not to associate pairs from the LibriSpeech and SAP datasets. In contrast, with a large τ, they are better harmonized, enhancing the robustness of the downstream regression model.*

should not be interpreted as clinical diagnoses. In addition, as dysarthric speech data are closely tied to health conditions, privacy protection and responsible data use are essential in any real-world deployment. Appropriate safeguards are necessary to ensure ethical use and to prevent misuse in sensitive decision-making contexts.

7. Conclusion and Future Works

In this work, we proposed a three-stage framework for robust dysarthric speech severity estimation that leverages unlabeled dysarthric speech and large-scale typical speech through pseudo-labeling and weakly supervised contrastive learning. By adapting Whisper representations toward task-relevant severity structure, our approach improves generalization across different etiologies, languages, and labeling schemes. Experimental results demonstrate strong cross-domain robustness, achieving an average SRCC of 0.761 on unseen datasets while maintaining in-domain performance; 0.415 and 0.311 point improvement compared to UTMOS and SpICE, respectively. Our ablation studies further show that weak supervision via coarse grouping in contrastive learning, along with a higher temperature (τ), plays a critical role in improving generalization.

In this work, training relied on a single dysarthric speech dataset, SAP. Although SAP is relatively large-scale, incorporating multilingual dysarthric datasets during training could further improve robustness. In addition, we showed that the learned representation space has the potential to serve as an interpretable severity dimension. Further analysis of this space to extract meaningful insights and understand the factors influencing predictions would be an important direction for future work. Finally, we plan to use the proposed severity predictor as an automatic labeling tool for unlabeled dysarthric speech and apply it to support downstream systems such as dysarthric ASR and speech enhancement.

8. Acknowledgments

This work was supported in part by the Institute of Information & communications Technology Planning & evaluation (IITP) grant funded by the Korean government (MSIT) [No. RS-2022-II220184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics] as well as the National Science Foundation under Grant No. 2512987.

9. Generative AI Use Disclosure

The authors acknowledge the use of an AI tool for copyediting and polishing the English language in this manuscript. The tool was used only to improve clarity, grammar, and style, and was not used to generate substantial portions of the manuscript or to develop the scientific content. All research design, experiments, analyses, and conclusions were conducted by the authors, who take full responsibility for the content of the paper.

10. References

[1] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021, pp. 6493–6497.

[2] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi *et al.*, "UTMOS: UTokyo-sarulab system for voicemos challenge 2022," in *Proc. Interspeech*, 2022, pp. 4521–4525.

[3] S. Venugopalan, J. Tobin, S. J. Yang, K. Seaver, R. J. Cave *et al.*, "Speech intelligibility classifiers from 550k disordered speech samples," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[4] R. L. MacDonald, P.-P. Jiang, J. Cattiau, R. Heywood, R. Cave *et al.*, "Disordered speech data collection: Lessons learned at 1 million utterances from project euphonia." in *Proc. Interspeech*, vol. 2021, 2021, pp. 4833–4837.

[5] J. Narain, V. Kowtha, C. Lea, L. Tooley, D. Yee *et al.*, "Voice Quality Dimensions as Interpretable Primitives for Speaking Style for Atypical Speech and Affect," in *Proc. Interspeech*, 2025, pp. 4628–4632.

[6] M. Hasegawa-Johnson, X. Zheng, H. Kim, C. Mendes, M. Dickinson *et al.*, "Community-supported shared infrastructure in support of speech accessibility," *Journal of Speech, Language, and Hearing Research*, vol. 67, no. 11, pp. 4162–4175, 2024.

[7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. of the International Conference on Machine Learning (ICML)*, vol. 119, 2020, pp. 1597–1607.

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 12 449–12 460.

[9] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov *et al.*, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[10] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[11] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey *et al.*, "Robust speech recognition via large-scale weak supervision," in *Proc. of the International Conference on Machine Learning (ICML)*, 2023, pp. 28 492–28 518.

[12] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian *et al.*, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 18 661–18 673.

[13] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang *et al.*, "Dysarthric speech database for universal access research," in *Proc. Interspeech*, 2008, pp. 1741–1744.

[14] H. Zhang, T. Zhang, G. Liu, D. Fu, X. Hou *et al.*, "DysArinVox: DYSphonia & DYSarthria mandARIN speech corpus," in *Proc. Interspeech*, 2024, pp. 932–936.

[15] R. Turrisi, A. Braccia, M. Emanuele, S. Giulietti, M. Pugliatti *et al.*, "EasyCall Corpus: A dysarthric speech dataset," in *Proc. Interspeech*, 2021, pp. 41–45.

[16] I. of Informatics of the Slovak Academy of Sciences, A. P. s.r.o., P.-E. University, M. Trnka, and M. Rusko, "EWA-DB – early warning of alzheimer speech database," 2023.

[17] J. Mendes-Laureano, J. A. Gómez-García, A. Guerrero-López, E. Luque-Buzo, J. D. Arias-Londoño *et al.*, "NeuroVoz: a castillian spanish corpus of parkinsonian speech," *Scientific Data*, vol. 11, no. 1, p. 1367, 2024.

[18] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 2001, pp. 749–752.

[19] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 626–630.

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. R. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 4214–4217.

[21] W. A. Jassim, J. Skoglund, M. Chinen, and A. Hines, "Warp-Q: Quality prediction for generative neural speech codecs," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021, pp. 401–405.

[22] W. C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda *et al.*, "The VoiceMOS Challenge 2022," in *Proc. Interspeech*, 2022, pp. 4536–4540.

[23] M. M. Halimeh, M. Torcoli, P. Grundhuber, and E. Habets, "On the relation between speech quality and quantized latent representations of neural codecs," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[24] A. Sanchez and S. King, "Can we reconstruct a dysarthric voice with the large speech model parler TTS?" in *Proc. Interspeech*, 2025, pp. 4138–4142.

[25] D. de Groot, T. Patel, D. Kayande, O. Scharenborg, and Z. Yue, "Objective and Subjective Evaluation of Diffusion-Based Speech Enhancement for Dysarthric Speech," in *Proc. Interspeech*, 2025, pp. 2740–2744.

[26] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey *et al.*, "Robust speech recognition via large-scale weak supervision," in *Proc. of the International Conference on Machine Learning (ICML)*, 2022.

[27] S.-W. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. Lai, K. Lakhotia *et al.*, "SUPERB: Speech processing universal performance benchmark," in *Proc. Interspeech*, 2021.

[28] Z. Ma, M. Chen, H. Zhang, Z. Zheng, W. Chen *et al.*, "EmoBox: Multilingual Multi-corpus Speech Emotion Recognition Toolkit and Benchmark," in *Proc. Interspeech*, 2024, pp. 1580–1584.

[29] K. Zha, P. Cao, J. Son, Y. Yang, and D. Katabi, "Rank-N-Contrast: Learning continuous representations for regression," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[30] J. Fan and D. S. Williamson, "JSQA: Speech quality assessment with perceptually-inspired contrastive pretraining based on jnd audio pairs," in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2025, pp. 1–5.

[31] M. Kim, J. Yoo, and H. Kim, "Dysarthric speech recognition using dysarthria-severity-dependent and speaker-adaptive models," in *Proc. Interspeech*, 2013.

[32] M. Soleymanpour, M. T. Johnson, R. Soleymanpour, and J. Berry, "Accurate synthesis of dysarthric speech for asr data augmentation," *Speech Commun.*, vol. 164, no. C, 2024.

[33] M. Merler, C. Agurto, J. Peller, E. Roitberg, A. Taitz *et al.*, "Clinical assessment and interpretation of dysarthria in ALS using attention based deep learning AI models," *NPJ Digital Medicine*, vol. 8, 2025.

[34] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[35] F. Wang and H. Liu, "Understanding the behaviour of contrastive loss," in *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2495–2504.

[36] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-invariance-covariance regularization for self-supervised learning," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2022.

[37] S. Team, "Silero VAD: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier," https://github.com/snakers4/silero-vad, 2024.

[38] S. Das, N. Singh, A. Gangwar, and S. Umesh, "Improved Intelligibility of Dysarthric Speech using Conditional Flow Matching," in *Proc. Interspeech*, 2025, pp. 2118–2122.

[39] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2017.

[40] L. van der Maaten and G. E. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.