DyPCL: Dynamic Phoneme-level Contrastive Learning for Dysarthric Speech Recognition

Wonjun Lee[*1], **Solee Im**[*2], **Heejin Do**[2],
Yunsu Kim[3], **Jungseul Ok**[1,2], **Gary Geunbae Lee**[1,2]
[1]Department of Computer Science and Engineering, POSTECH, Republic of Korea
[2]Graduate School of Artificial Intelligence, POSTECH, Republic of Korea
[3]aiXplain Inc., Los Gatos, CA, USA
{lee1jun, solee0022, heejindo, jungseul.ok, gblee}@postech.ac.kr, yunsu.kim@aixplain.com

Abstract

Dysarthric speech recognition often suffers from performance degradation due to the intrinsic diversity of dysarthric severity and extrinsic disparity from normal speech. To bridge these gaps, we propose a Dynamic Phoneme-level Contrastive Learning (DyPCL) method, which leads to obtaining invariant representations across diverse speakers. We decompose the speech utterance into phoneme segments for phoneme-level contrastive learning, leveraging dynamic connectionist temporal classification alignment. Unlike prior studies focusing on utterance-level embeddings, our granular learning allows discrimination of subtle parts of speech. In addition, we introduce dynamic curriculum learning, which progressively transitions from easy negative samples to difficult-to-distinguish negative samples based on phonetic similarity of the phoneme. Our approach to training by difficulty levels alleviates the inherent variability of speakers, better identifying challenging speeches. Evaluated on the UASpeech dataset, DyPCL outperforms baseline models, achieving an average 22.10% relative reduction in word error rate (WER) across the overall dysarthria group.

1 Introduction

Accurate recognition of dysarthric speech, which is slurred and difficult to understand, is critical for assisting effective communication for individuals with speech impairments (Young, 2010). However, due to the inherent diversity of severity levels and substantial differences compared to normal speech, dysarthric speech recognition (DSR) poses significant challenges. Previous studies mainly focused on data augmentation (Prananta et al., 2022; Jiao et al., 2018; Wang et al., 2023a) and speaker-adaptive training (Yu et al., 2018; Hu et al., 2019; Lin et al., 2024). However, relying on additional augmentation techniques or feature extraction methods limits the practical applicability.

Contrastive learning has been explored in DSR to learn invariant representations by using healthy speech as a stable reference point (Wu et al., 2021; Wang et al., 2024b). The model can more effectively capture the underlying linguistic content in dysarthric speech by anchoring the learning process on phonetic embeddings from healthy speakers despite surface-level variations. For instance, Wu et al. (2021) applied pyramid pooling to distinguish words within audio, while Wang et al. (2024b) used word-level contrastive learning with entire audio segments. However, word-level embeddings fail to achieve fine-grained recognition, which is crucial for dysarthric speakers with distinct pronunciation challenges.

In this paper, we propose a dynamic phoneme-level contrastive learning (DyPCL) framework, integrating phoneme-level speech embedding with contrastive learning. DyPCL incorporates two-way dynamic approaches: dynamic connectionist temporal classification (CTC) alignment and dynamic curriculum learning. First, we introduce a dynamic CTC alignment method that accurately aligns speech embeddings with phoneme labels for phoneme-level contrastive learning. Unlike previous approaches that rely on external alignment modules for phoneme-level contrastive learning (Fu et al., 2022), dynamic CTC alignment simultaneously learns robust feature representations. It aligns speech sequences with their corresponding phonemes during training, eliminating the need for explicit frame-level annotations.

In addition, we introduce dynamic curriculum learning, which dynamically organizes negative samples based on difficulty, determined by the similarity distance of the anchor and negative phoneme in PCL. This phonetic approach further enhances DSR performance by effectively distinguishing between similar-sounding phonemes, a critical factor

[*]Equally contributed

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4701–4712
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

in dysarthric speech.

Evaluated on the UASpeech dataset, a representative DSR benchmark, our method shows substantial improvements over baseline models. Specifically, in the lowest intelligibility group, DyPCL reduces the word error rate (WER) from 58.49% to 49.45%, while overall WER across all dysarthria groups drops from 25.97% to 20.23%. Extensive analysis and ablation studies highlight the robustness of the proposed strategies.

2 Related Work

Dysarthric Speech Recognition Prior studies have primarily utilized data augmentation and speaker-adaptive training to address DSR challenges. Augmentation methods like speed and temporal perturbation (Prananta et al., 2022; Geng et al., 2022) simulate dysarthric speech characteristics, while adversarial training (Jiao et al., 2018; Huang et al., 2022; Jin et al., 2023b,a; Wang et al., 2024a) and diffusion models (Wang et al., 2023a) synthesize dysarthric speech. Speaker-adaptive training helps models handle speaker variability through features like Learning Hidden Unit Contributions (Yu et al., 2018; Geng et al., 2023b,a), x-vector (Baskar et al., 2022), and Acoustic-to-Articulatory inversion models (Hu et al., 2019; Liu et al., 2020; Hu et al., 2022, 2024; Hsieh and Wu, 2024a; Lin et al., 2024; Hu et al., 2023). However, these methods require additional datasets and external models, which increases computational complexity. We simplify the DSR process by using only the UASpeech dataset and a single ASR model, eliminating the need for external resources.

Contrastive Learning Contrastive learning in speech recognition has proven effective in various atypical scenarios, such as noisy environments (Wang et al., 2022; Zhu et al., 2023) and accented speech datasets (Han et al., 2021; Fu et al., 2022), due to its ability to learn robust speech representations by reducing speech embedding variability. For DSR, contrastive learning has been applied to reduce the distance between healthy utterance and dysarthric utterance representations (Wu et al., 2021; Wang et al., 2024a) using the word or utterance fragments. To consider subtle segments during training, phoneme-level contrastive learning has been suggested for accented speech recognition (Fu et al., 2022), but it has yet to be explored in DSR.

To perform phoneme-level contrastive learning, sophisticated phoneme alignment for a given user utterance is required. Previous research explored various alignment methods (Rousso et al., 2024; Gorman et al., 2011; McAuliffe et al., 2017) to match given phonemes to audio frames closely. Notably, using CTC forced alignment has shown great alignment accuracy (Huang et al., 2024; Zhao and Bell, 2024). This work proposes phoneme-level dynamic CTC alignment that dynamically adapts during optimizing contrastive learning. In addition, considering the importance of hard negative sampling for contrastive learning (Robinson et al., 2021; Kalantidis et al., 2020), we dynamically select negative samples across varying difficulty groups and phoneme distance levels, integrating with the curriculum learning process; thus, the model can learn invariant representations.

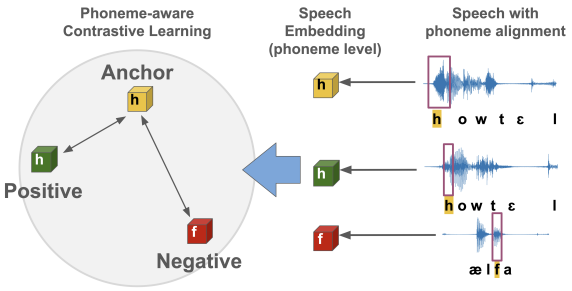

Figure 1: Phoneme-level Contrastive Learning. The phoneme-aligned speech segment corresponding to the phoneme "h" in the word "HOTEL" is used as both the anchor and the positive sample, while the phoneme "f" from the word "ALPHA" serves as the negative sample.

3 DyPCL

3.1 Phoneme-level Contrastive Learning

To conduct DyPCL, we leverage phoneme-level contrastive learning (PCL) with CTC loss in a multitask learning framework. PCL is a training strategy designed to learn phoneme-level representations for speech recognition through contrastive learning (Figure 1). PCL effectively clusters and separates targeted phoneme embeddings by focusing on audio segments corresponding to phonemes. This approach is particularly robust for tasks involving single-word audio, where even a minor phoneme error can significantly impact intelligibility. Further, PCL benefits CTC models that use the same phonemes as output units by enhancing the model's ability to distinguish subtle phonetic variations, thereby improving recognition accuracy.

We design two training stages: (1) CTC train-

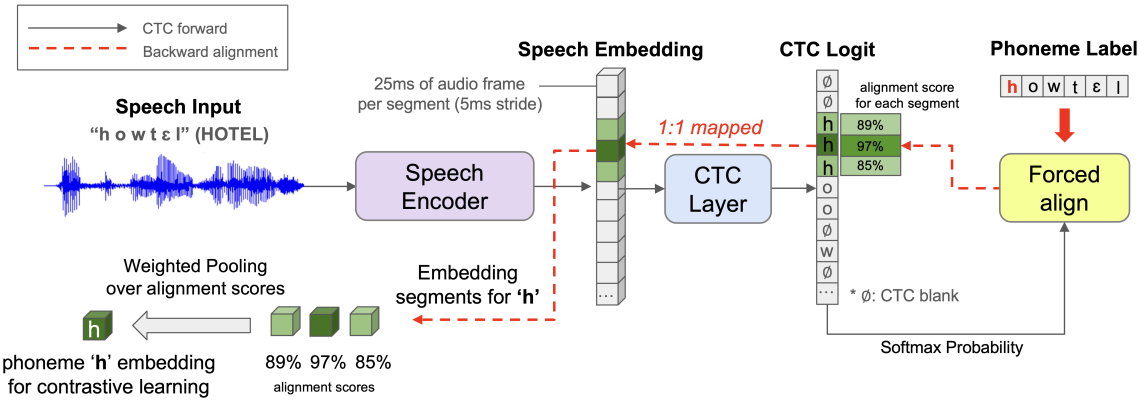

Figure 2: Phoneme embedding extraction for the target phoneme "h" in the word "HOTEL" using **Dynamic CTC Alignment** for phoneme-level contrastive learning.

ing and (2) combined CTC and PCL training for contrastive learning. The first stage, CTC training, targets phoneme recognition based on given speech and labels, following procedures in general speech recognition. In the second stage, CTC/PCL training integrates phoneme-level contrastive learning specifically for dysarthric speech while maintaining CTC training for phoneme recognition.

We pair anchor, positive, and negative samples to construct the contrastive learning dataset from UASpecch (Kim et al., 2023). We decompose words into phonemes using the phonemizer tool (Mortensen et al., 2018). Anchor samples are exclusively selected from the control (C) group, healthy speech, to serve as a reference for correct pronunciation. Positive samples are the same word and phoneme as the anchor but taken from dysarthric groups (H, M, L, VL) (refer to 4.2). Negative samples comprise different words and phonemes from the anchor and positive, also drawn from dysarthric groups to introduce challenging contrasts. For the healthy-speech-only trainset (B2-Control of UASpeech), positive samples are selected from different speakers within the same group.

This process yields roughly 48.65 billion triplet pairs. We use stratified sampling to balance training cost and efficiency, limiting each anchor to a maximum of five positive samples and each anchor-positive pair to five negative samples. Consequently, 1.18 million triplet pairs are created for contrastive learning. During training, we randomly sampled 200,000 triplet pairs.

The triplet loss works by ensuring that the embedding of an anchor sample a is closer to a positive sample p than to a negative sample n, by at least a given margin. The triplet loss, $L_{\text{triplet}}(a, p, n)$, with margin m is defined as:

$$\max\left(0,\ \|f(\mathbf{a}) - f(\mathbf{p})\|_2^2 - \|f(\mathbf{a}) - f(\mathbf{n})\|_2^2 + m\right) \quad (1)$$

Here, $f(x)$ represents the speech embedding for phoneme x obtained from the speech encoder, where $f(\mathbf{a})$ is the embedding of the anchor, $f(\mathbf{p})$ is the embedding of the positive sample, and $f(\mathbf{n})$ is the embedding of the negative sample. The squared Euclidean distances $\|\cdot\|_2^2$ measure how far apart these embeddings are. The margin ensures the anchor is closer to the positive than the negative by a certain threshold, encouraging the model to learn distinct representations for different phonemes. Our total multitask loss function for CTC/PCL training, L_{total}, is defined as follows:

$$\left(\frac{1}{3}\sum_{i}^{(A,P,N)} L_{\text{CTC}}(i)\right) + \lambda \cdot L_{\text{triplet}}(a, p, n) \quad (2)$$

where A, P, and N denote the anchor, positive, and negative audio samples, respectively, and a, p, and n are the corresponding anchor, positive, and negative phonemes. In this work, we set $\lambda = 0.5$. The following subsections outline the dynamic components of PCL, forming the basis of DyPCL.

3.2 Dynamic CTC Alignment

To perform PCL, we need a phoneme-level alignment for each audio sample. This requires accurately mapping each phoneme to its corresponding speech embedding. Fu et al. (2022) demonstrate that phoneme-level contrastive learning can improve ASR accuracy in accented speech. They utilize an HMM-DNN acoustic model for forced

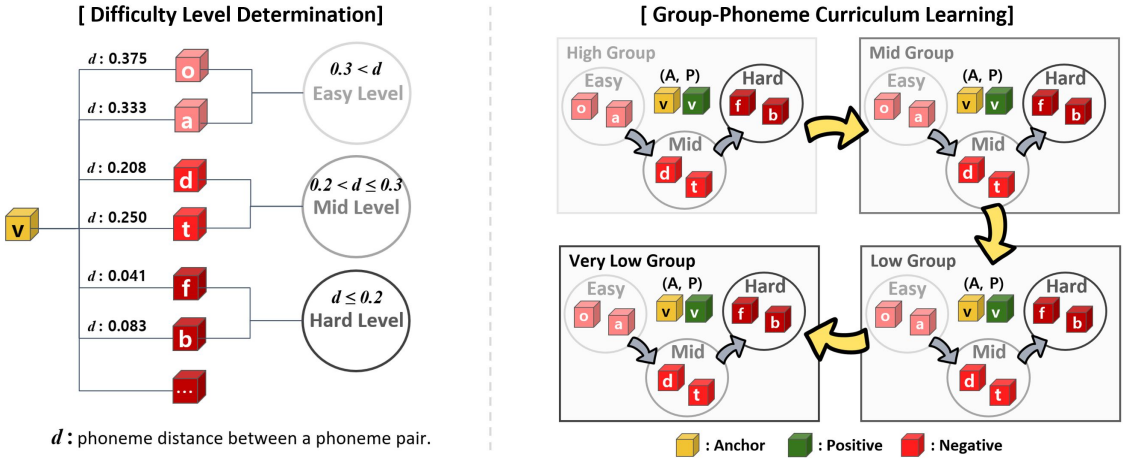

Figure 3: The illustration of difficulty level determination in three (easy, mid, hard) levels by phoneme distance measurement (Left) and the group-phoneme (GP) curriculum learning (Right). The figure shows an example where the anchor and positive samples are "v".

alignment using the Kaldi toolkit (Povey et al., 2011). Although the authors acknowledge that the alignment may not be perfect, their results show substantial improvements with phoneme-level contrastive learning.

In contrast, our research focuses on dysarthric speech recognition, where pre-trained forced alignment models often struggle to accurately align speech with labels, particularly for groups with low intelligibility. This misalignment arises because these models are typically trained on standard speech datasets, while the acoustic properties of dysarthric speech vary significantly from typical speech. Such discrepancies in alignment can critically impair phoneme-level contrastive learning.

Forced alignment is conventionally done by timestamping (Rousso et al., 2024), which provides phoneme boundaries about audio frames. However, for our purposes, we need a method that pinpoints the corresponding speech embedding for each phoneme within a CTC model; thus, a direct solution is required to eliminate unnecessary errors. Figure 2 presents an overview of the proposed dynamic CTC alignment. Drawing from CTC forced alignment[1], we directly extract speech embeddings for specific phonemes. CTC forced alignment maps audio to transcription by predicting the most likely alignment between speech frames and text based on CTC logits, handling timing variations.

The output provides alignment scores for each CTC logit. Given the speech embeddings shaped $[\textit{embedding size}, \textit{sequence length}]$ and CTC logits shaped $[\textit{CTC vocabulary size}, \textit{sequence length}]$, we can map the alignment scores to the corresponding speech embeddings for each phoneme, as illustrated by the red lines (backward alignment) in Figure 2. As the speech encoder (wav2vec2.0 (Baevski et al., 2020) and its variants) generates one embedding token per 25 ms of audio, multiple indices in both the embedding and logits can map to a single phoneme. We generate a single phoneme representation using weighted pooling with alignment scores as weights. The weighted pooled phoneme embeddings for each anchor ($f(\mathbf{a})$), positive ($f(\mathbf{p})$), and negative ($f(\mathbf{n})$) samples will be used in the triplet loss, as defined in Equation 1. The speech encoder and CTC layer will be updated during CTC/PCL training, leading to more accurate alignments, and these improved alignments will further enhance the training process.

3.3 Dynamic Curriculum Learning

Negative sampling in contrastive learning is crucial, as selecting hard negatives can significantly enhance model performance (Robinson et al., 2021; Kalantidis et al., 2020; Srinidhi and Martel, 2021). Each phoneme is treated as an anchor, positive, or negative sample in PCL. Anchors are selected from the control group (C), serving as the reference, while positives are chosen from the same word utterance as the anchor but from the dysarthric group (H, M, L and VL. Refer to Table 5). Negatives are randomly sampled from other phonemes within the dysarthric group, which do not directly relate to the anchor or positive.

In the DyPCL framework, we dynamically se-

[1] https://pytorch.org/audio/main/generated/torchaudio.functional.forced_align.html

lect negative samples using a curriculum learning approach. This dynamic selection strategy gradually increases the difficulty of negative samples, fostering a more robust learning process. The difficulty is determined by phoneme distance, which is measured using the PanPhon tool (Mortensen et al., 2016) [2]. The *hamming feature edit distance* is used to calculate phoneme distance, with equal weighting across all 21 articulatory features that define each phoneme. The phoneme distance reflects how similar or different phonemes sound phonetically, ranging from 0.0416 (most similar) to 0.583 (most different). Figure 6 in Appendix A shows the phoneme distance matrix in a heat map. By learning to distinguish similar-sounding phonemes in the embedding space through DyPCL, the model can further improve recognition accuracy for given phonemes.

We differentiate the curriculum by varying the phoneme distance of negative samples in multiple ways. As shown on the left side of Figure 3, we categorize the difficulty into three levels: easy ($d > 0.3$), medium ($0.2 < d \leq 0.3$), and hard ($d \leq 0.2$), where d represents the phoneme distance between an anchor and a negative sample. Further difficulty variations are discussed in Section 5.3.

To further optimize the effectiveness of our phoneme distance-based curriculum (**P**), we incorporate a group-level curriculum (**G**) as suggested in Hsieh and Wu (2024b). This method trains the DSR model progressively, following an intelligibility group order from H to M, L, and VL, which has improved DSR accuracy compared to non-ordered training. We enhance the effectiveness of the curriculum by combining both P and G strategies.

The resulting **GP** (group first, then phoneme distance) curriculum first organizes the groups in the H, M, L, and VL order and then applies the P strategy within each group. The GP curriculum comprises 12 levels (4 groups $\times$ 3 phoneme difficulty levels), as illustrated on the right side of Figure 3. The **PG** (phoneme distance first, then group) curriculum is also evaluated in Section 5.1. All curricula were designed to train on 200,000 triplet pairs per epoch.

4 Experiment Setup

4.1 Model & Training

In our experiments, we utilize a CTC head with several pretrained speech encoders, including Wav2Vec2.0[3](Baevski et al., 2020), HuBert[4](Hsu et al., 2021), and WavLM[5](Chen et al., 2022), each with 315M parameters and a CTC head of 44K parameters. All speech encoders have an embedding size of 1024. We extract phoneme-level speech embeddings $f(x)$ sized [1024, 1] via dynamic CTC alignment for PCL.

We first trained the model using only CTC loss in the initial stage, followed by combined CTC/PCL training. This two-stage approach helped refine the dynamic CTC alignment for optimal performance in PCL training. The model was optimized using the AdamW (Loshchilov and Hutter, 2017) algorithm with parameters $(\beta_1, \beta_2) = (0.9, 0.99)$, a learning rate of 3×10^{-4}, weight decay of 1×10^{-5}, and batch sizes of 128 for CTC training and 64 for CTC/PCL training. We trained the models on 8 NVIDIA A100 GPUs, utilizing a linear learning rate scheduler and selecting the best model based on the lowest overall WER on the validation set. The CTC training ran for 100 epochs (10 hours), and the CTC/PCL stage ran for 5 epochs (20 hours).

4.2 Dataset & Preprocessing

The recent release of UASpeech (Kim et al., 2023) includes 13 speakers in a control group (denoted as C) and 15 speakers with dysarthria, categorized into intelligibility levels: High (H), Mid (M), Low (L), and Very Low (VL). For more details, please refer to Table 5 in Appendix A.

We follow prior works for training split (**TRAIN**) (Hu et al., 2022; Geng et al., 2023a; Hu et al., 2024; Hsieh and Wu, 2024b) which has all audio from B1, B3 and B2 of control group (B2-Control). For test splits, we use two sets: B2 of all dysarthria groups (**TEST**) (Geng et al., 2023a; Hu et al., 2024; Hsieh and Wu, 2024b), and only the common words (CW) of dysarthria group excluding uncommon words (UW) (**CTEST**) (Bhat et al., 2022b; Wang et al., 2024b). We randomly sampled 10% of the TRAIN set for validation split. Every audio recording is from microphone 5 (M5) in UASpeech.

[2] https://github.com/dmort27/panphon

[3] https://huggingface.co/facebook/wav2vec2-large-960h

[4] https://huggingface.co/facebook/hubert-large-ls960-ft

[5] https://huggingface.co/microsoft/wavlm-large

Configuration				UASpeech WER (%)				
Speech Encoder	Loss Function	CL Target	Neg. Sampling	H	M	L	VL	ALL
Wav2Vec2.0	CTC	-	-	6.54	21.55	31.89	61.04	29.15
WavLM	CTC	-	-	5.32	20.10	26.43	58.94	26.80
HuBERT ✓	CTC	-	-	**4.33**	**19.32**	**25.32**	**58.49**	**25.97**
	CTC+CL	word	R	5.84	18.63	24.40	58.50	26.15
	CTC+CL	phoneme	R	**3.12**	**15.21**	**21.16**	**53.72**	**22.64**
HuBERT ✓	CTC+CL	phoneme	G	3.24	15.43	20.81	50.77	21.87
	CTC+CL	phoneme	P	2.75	14.56	18.45	51.26	21.19
	CTC+CL	phoneme	PG	**2.73**	13.21	18.20	50.21	20.58
	CTC+CL	phoneme	GP	2.77	**12.98**	**17.60**	**49.45**	**20.23**

Table 1: WER on UASpeech **TEST** set with different configurations: Speech Encoder, Loss function, Contrastive Learning (CL) target, and Negative sampling method. **R** in negative sampling represent random sampling and **G**, **P**, **PG** and **GP** represent corresponding curriculum strategy described in Section 3.3. "ALL" denotes the average WER across four groups, weighted by the number of speakers in each group.

5 Result & Analysis

5.1 Main Result

In Table 1, we first evaluate the performance of three speech encoders—Wav2Vec2.0, WavLM, and HuBERT—using CTC training. As indicated in several studies (Wang et al., 2024b; Hu et al., 2024), the HuBERT model achieves the best WER across all speaker groups and the overall average. Consequently, we conducted further experiments using the HuBERT model.

To assess the impact of phoneme-level contrastive learning, we trained the HuBERT-CTC model using word- and phoneme-level contrastive learning. Since the UASpeech dataset consists of isolated word recordings, word-level alignment was not required for word-level contrastive learning. Our results indicate that word-level contrastive learning underperformed, even falling short of the baseline CTC model. In contrast, phoneme-level contrastive learning significantly reduced WER across all speaker groups and the overall average. Notably, the VL group showed a marked improvement, with WER decreasing from **58.49%** using CTC alone to **53.72%** with PCL.

Then, we evaluate the negative sampling strategies within PCL. All curriculums (G, P, PG, GP) show improvements compared to the PCL model with random negative sampling (R). When comparing group-level (G) and phoneme-level (P) curricula, we found that P achieved better overall performance (21.18% vs. 21.87%), though G performed better in the VL group (50.77% vs. 51.26%). We attribute this to the fact that the VL group is trained last in the G curriculum.

When combining both strategies, **PG** and **GP** yielded substantial improvements over the individual methods. In particular, the **GP** curriculum achieved the best overall WER of **20.23%** and **49.45%** in VL, demonstrating that curriculum learning with a sophisticated difficulty progression can further enhance DSR performance.

Phoneme alignment method	UASpeech WER(%)				
	H	M	L	VL	ALL
CTC forced align (timestamp)	3.67	19.21	26.49	58.71	26.02
CTC forced align (logit level)	3.65	18.32	26.91	57.43	25.58
Dynamic CTC alignment	**3.12**	**15.21**	**21.16**	**53.72**	**22.64**

Table 2: Effect of different alignment methods for PCL. Curriculum learning is not applied in this evaluation (random sampling is used) to isolate the impact of the alignment methods.

5.2 Effect of Dynamic CTC Alignment

As discussed in Section 3.2, conventional phoneme alignment models struggle to accurately align phonemes in dysarthric speech, particularly for speakers with low intelligibility. This misalignment can severely impact the effectiveness of PCL, as it relies on extracting precise phoneme embeddings. To address this, we evaluate the effectiveness of dynamic CTC alignment by comparing it to conventional alignment methods.

For a fair comparison, we use the HuBERT-CTC model trained on the **TRAIN** set as a baseline for CTC forced alignment (Table 1). Phoneme alignment can be implemented in two primary ways: (1) using the timestamps of target phonemes in the au-

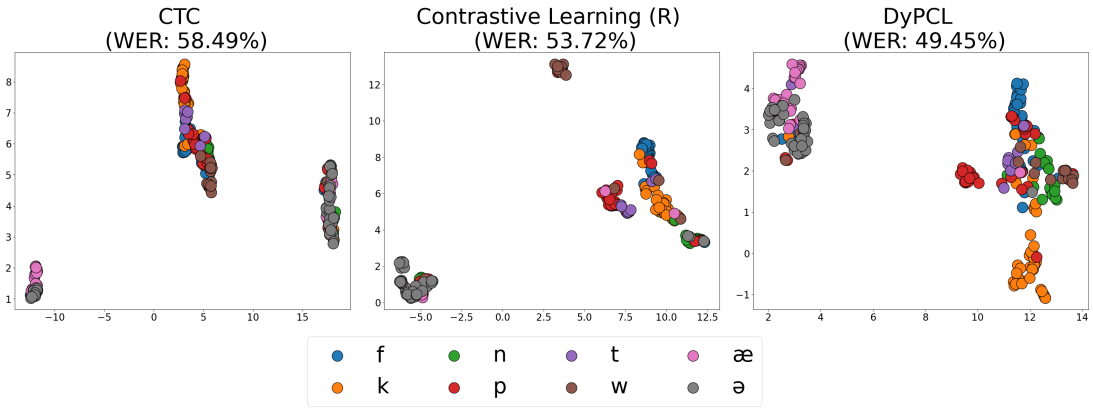

Figure 4: UMAP Visualization of Phoneme embeddings on **TEST** set (**VL** Group only): Phoneme embeddings, extracted via forced CTC alignment (Figure 2), are shown for three models: CTC, Contrastive learning with random sampling (**R**), and DyPCL with group-phoneme level curriculum (**GP**). Points are color-coded by phoneme, illustrating how each model clusters and separates them. For each phoneme, up to 100 embeddings were displayed.

dio to extract phoneme embeddings by calculating their corresponding embedding indices based on a 25ms window with a 5ms stride (Baevski et al., 2020), and (2) applying backward alignment on CTC logits, as proposed in the dynamic CTC alignment. The key difference between the second approach and dynamic CTC alignment is that the alignment model in the former is not updated during training.

In Table 2, the use of alignment with timestamps yielded underwhelming results, showing only marginal improvements for the H and M groups compared to the HuBERT-CTC model results in Table 1. This result can be attributed to incorrect alignments and potential conversion errors between timestamps and phoneme embeddings. When alignment was applied at the logit level, we achieved overall improvements, though there was some degradation in the VL group. In contrast, dynamic CTC alignment substantially improved WER across all groups. These gains are attributed to the model being optimized with both CTC and PCL losses, which enhance alignment accuracy and, in turn, lead to better DSR accuracy.

5.3 Analysis on Curriculum Difficulty Levels

In Figure 5, the average and median phoneme distances are 0.28 and 0.29, respectively. Based on this, we initially set the threshold at 0.3 to divide the difficulty into two levels: easy and hard (2 LV). We then refined the division by adding a mid-difficulty level at 0.2, creating three levels: easy, mid, and hard (3 LV). Furthermore, we explored finer granularity by dividing the phoneme distance range into 0.1 intervals, resulting in six difficulty levels (6 LV).

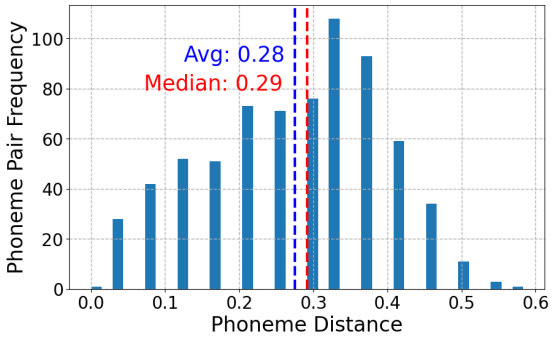

Figure 5: Distribution of Phoneme distance over phoneme pairs.

Phoneme distance ranges	UASpeech WER(%)				
	H	M	L	VL	ALL
(2 LV) $0 < H \leq 0.3 < E \leq 0.583$	**2.75**	13.25	17.94	50.43	20.6
(3 LV) $0 < H \leq 0.2 < M \leq 0.3 < E \leq 0.583$	2.77	**12.98**	**17.6**	**49.45**	**20.23**
(6 LV) divide every 0.1 distance	2.94	13.70	18.03	51.43	21.04

Table 3: Difficulty levels for phoneme distance curriculum. WER is evaluated on **TEST**.

Table 3 illustrates how difficulty levels in dynamic curriculum learning were established by segmenting the range of phoneme distances. Our results show that dividing the phoneme distance into three levels (easy, mid, and hard) yielded the best performance for our DyPCL. Although a two-level division produced comparable results, the three-level split outperformed it overall. However, the six-level division, which introduces extremely easy ($d \leq 0.1$) and extremely hard ($0.5 \leq d$) negative pairs, led to suboptimal results. This comparison highlights the trade-offs between granularity and performance, as the narrow distance intervals often resulted in some phonemes lacking suitable nega-

tive pairs, which adversely affected performance in our evaluations. Figure 7 in Appendix A shows the distribution of phoneme pairs across the three difficulty levels (3 LV) for each phoneme.

5.4 Phoneme Embedding Discrimination and Clustering

Figure 4 presents the Uniform Manifold Approximation and Projection (McInnes et al., 2018) (UMAP) visualization of phoneme embeddings for the **TEST** VL group across three models. The distribution of phoneme embeddings illustrates how well each model distinguishes and clusters the phonemes. In the CTC model, phoneme embeddings are not well-separated and form small, unclear clusters. The contrastive learning model shows more distinct clustering, although some phonemes still appear ambiguously grouped. In contrast, the DyPCL model demonstrates a clear and decisive separation of phonemes. Notably, even very similar-sounding phonemes, such as "**æ**" and "**ə**" (marked in pink and grey, respectively) with a phoneme distance of 0.125, are well-separated in DyPCL. This improvement is attributed to the informative curriculum learning strategy, which progressively trains the model to better distinguish similar phonemes.

5.5 Comparison with Benchmarks

Table 4 presents a comparison of the performance of our model against state-of-the-art (SOTA) methods on the **CTEST** set. Bhat et al. (2022a) employs a two-stage augmentation approach. The second and third methods come from Wang et al. (2024b), which was the first to introduce contrastive learning for UASpeech. In their study, the "Speaker Dependent (SD) w/ finetune" method focuses on fine-tuning speaker-specific word prototypes, while the "Speaker Independent (SI)" method works across both word- and speaker-level instances to improve generalization. The SD approach improved the WER to 13.49%, while the SI method further reduced it to 12.09%. In contrast, our model, DyPCL (GP), achieved the lowest WER at severity levels, significantly reducing the overall WER to 10.34%, outperforming all previous models.

Table 4 also compares our model, DyPCL (GP), against other methods on the **TEST** set. For a fair comparison, the results from previous works are reported without data augmentation (DA), focusing on the core contributions of each method. Wang et al. (2023b) used hyperparameter adaptation to handle speaker differences, achieving an overall WER of 30.49%. Hsieh and Wu (2024a) applied curriculum learning, training progressively from high to low intelligibility groups with their proposed re-grouping method, but still reported a relatively high WER of 7.99% for the easiest group (H), highlighting challenges in early-stage learning. Both Hu et al. (2023) and Hu et al. (2024) used speaker-adaptive training, incorporating speaker-specific articulatory and acoustic features, achieving WERs of 24.62% and 22.62%, respectively. Geng et al. (2023a) integrated severity information and system combination, and without DA, reported an overall WER of 24.57%.

Model	UASpeech WER(%)				
	H	M	L	VL	ALL*
CTEST					
Bhat et al. (2022a)	6.40	14.6	18.9	61.50	25.35
Wang et al. (2024b) (SD w/ finetune)	5.12	4.89	6.27	37.67	13.49
Wang et al. (2024b) (SI)	2.35	6.01	7.91	32.11	12.09
DyPCL (GP)	**1.09**	**3.94**	**5.02**	**31.33**	**10.34**
TEST					
Wang et al. (2023b)	5.22	21.35	33.37	62.04	30.49
Hsieh and Wu (2024a)	7.99	16.12	22.28	52.15	24.64
Hu et al. (2023)	6.32	14.04	25.03	53.12	24.62
Geng et al. (2023a) (w/o DA)	2.91	12.10	23.91	59.38	24.57
Hu et al. (2024)	4.20	**12.06**	23.51	50.7	22.62
DyPCL (GP)	**2.77**	12.98	**17.6**	**49.45**	**20.7**

Table 4: WER comparison on the **CTEST** and **TEST**, showing the performance of DyPCL (GP) against previous studies. *: unweighted average over groups.

Our model, DyPCL (GP), not only achieved the lowest overall WER of 20.7%, outperforming all previous methods but also demonstrated significant improvements for the low intelligibility groups. With WERs of 17.6% for the L group and 49.45% for the VL group, it maintained strong performance across all dysarthria severity levels, achieving a WER of just 2.77% for the High (H) intelligibility group. This level of robustness highlights the reliability of its performance.

In Table 4, "ALL*" represents the recalculated unweighted average WER across the four dysarthria groups, ensuring consistency and fairness in our comparison process. It is important to note that variations in reported WERs in different papers may arise from additional factors, such as the inclusion of control groups, which we have taken into account.

6 Conclusion

This paper introduced the Dynamic Phoneme-level Contrastive Learning (DyPCL) framework to im-

prove dysarthric speech recognition. DyPCL effectively tackles phoneme alignment challenges and accounts for phonetic difficulty through dynamic CTC alignment and curriculum learning. Our experiments on the UASpeech dataset demonstrated the effectiveness of DyPCL, which reduced the WER from 58.49% to 49.45% in the Very Low (VL) intelligibility group and the overall WER across all dysarthria groups from 25.97% to 20.23%. These results underscore DyPCL's capability to capture subtle phonetic variations, significantly enhancing speech recognition accuracy across all levels of dysarthria severity

Limitations

While DyPCL has shown strong performance in recognizing dysarthric speech, its reliance on paired data, such as in the UASpeech dataset, where each dysarthric speech sample is paired with a corresponding control group utterance, suggests that there may be opportunities to further generalize the model. This pairing provides valuable reference points for contrastive learning, but by focusing on phoneme embeddings rather than specific word pairs, future research could explore the model's applicability in scenarios where such paired data is unavailable. This shift could potentially broaden DyPCL's utility in more diverse environments where only unpaired or less structured data is available.

Moreover, we did not employ data augmentation (DA) techniques in this study to ensure a clear evaluation of DyPCL's core contributions. However, given that previous research indicates the positive impact of DA on dysarthric speech recognition, combining DyPCL with DA strategies could yield further improvements. Future work will explore these possibilities to enhance performance and generalizability.

Acknowledgments

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2019-II191906, Artificial Intelligence Graduate School Program(POSTECH)) (5%) and was supported by the IITP(Institute of Information & Coummunications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2025-RS-2024-00437866)(47.5%) and was supported by Smart HealthCare Program funded by the Korean National Police Agency(KNPA) (No. 220222M01) (47.5%)

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Murali Karthick Baskar, Tim Herzig, Diana Nguyen, Mireia Diez, Tim Polzehl, Lukas Burget, and Jan Černocký. 2022. Speaker adaptation for wav2vec2 based dysarthric asr. In *Proceedings of Interspeech 2022*, pages 3403–3407.

C. Bhat, A. Panda, and H. Strik. 2022a. Improved ASR performance for dysarthric speech using two-stage data augmentation. In *Proceedings of Interspeech 2022*, pages 46–50.

Chitralekha Bhat, Ashish Panda, and Helmer Strik. 2022b. Improved asr performance for dysarthric speech using two-stage dataaugmentation. In *INTERSPEECH*, pages 46–50.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Li Fu, Xiaoxiao Li, Runyu Wang, Lu Fan, Zhengchen Zhang, Meng Chen, Youzheng Wu, and Xiaodong He. 2022. Scala: Supervised contrastive learning for end-to-end speech recognition. In *Proceedings of Interspeech 2022*.

Mengzhe Geng, Zengrui Jin, Tianzi Wang, Shujie Hu, Jiajun Deng, Mingyu Cui, Guinan Li, Jianwei Yu, Xurong Xie, and Xunying Liu. 2023a. Use of speech impairment severity for dysarthric speech recognition. In *Proceedings of Interspeech 2023*, pages 2328–2332.

Mengzhe Geng, Xurong Xie, Rongfeng Su, Jianwei Yu, Zengrui Jin, Tianzi Wang, Shujie Hu, Zi Ye, Helen Meng, and Xunying Liu. 2023b. On-the-fly feature based rapid speaker adaptation for dysarthric and elderly speech recognition. In *Proceedings of Interspeech 2023*, pages 1753–1757.

Mengzhe Geng, Xurong Xie, Zi Ye, Tianzi Wang, Guinan Li, Shujie Hu, Xunying Liu, and Helen Meng. 2022. Speaker adaptation using spectro-temporal deep features for dysarthric and elderly speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2597–2611.

Kyle Gorman, Jonathan Howell, and Michael Wagner. 2011. Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian acoustics*, 39(3):192–193.

Tao Han, Hantao Huang, Ziang Yang, and Wei Han. 2021. Supervised contrastive learning for accented speech recognition. *arXiv preprint arXiv:2107.00921*.

I-Ting Hsieh and Chung-Hsien Wu. 2024a. Dysarthric speech recognition using curriculum learning and articulatory feature embedding. In *Interspeech 2024*, Kos, Greece. ISCA.

I-Ting Hsieh and Chung-Hsien Wu. 2024b. Dysarthric speech recognition using curriculum learning and articulatory feature embedding. In *Proceedings of Interspeech 2024*, pages 1300–1304, Kos, Greece. International Speech Communication Association (ISCA).

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Shoukang Hu, Shansong Liu, Heng Fai Chang, Mengzhe Geng, Jiani Chen, Lau Wing Chung, To Ka Hei, Jianwei Yu, Ka Ho Wong, Xunying Liu, et al. 2019. The cuhk dysarthric speech recognition systems for english and cantonese. In *INTERSPEECH*, pages 3669–3670.

Shujie Hu, Shansong Liu, Xurong Xie, Mengzhe Geng, Tianzi Wang, Shoukang Hu, Mingyu Cui, Xunying Liu, and Helen Meng. 2022. Exploiting cross domain acoustic-to-articulatory inverted features for disordered speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6747–6751. IEEE.

Shujie Hu, Xurong Xie, Mengzhe Geng, Zengrui Jin, Jiajun Deng, Guinan Li, Yi Wang, Mingyu Cui, Tianzi Wang, Helen Meng, et al. 2024. Self-supervised asr models and features for dysarthric and elderly speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Shujie Hu, Xurong Xie, Zengrui Jin, Mengzhe Geng, Yi Wang, Mingyu Cui, Jiajun Deng, Xunying Liu, and Helen Meng. 2023. Exploring self-supervised pre-trained asr models for dysarthric and elderly speech recognition. In *Proceedings of ICASSP 2023*, pages 1–5.

Ruizhe Huang, Xiaohui Zhang, Zhaoheng Ni, Li Sun, Moto Hira, Jeff Hwang, Vimal Manohar, Vineel Pratap, Matthew Wiesner, Shinji Watanabe, et al. 2024. Less peaky and more accurate ctc forced alignment by label priors. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11831–11835. IEEE.

Wen-Chin Huang, Bence Mark Halpern, Lester Phillip Violeta, Odette Scharenborg, and Tomoki Toda. 2022. Towards identity preserving normal to dysarthric voice conversion. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6672–6676. IEEE.

Yishan Jiao, Ming Tu, Visar Berisha, and Julie Liss. 2018. Simulating dysarthric speech for training data augmentation in clinical speech applications. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6009–6013. IEEE.

Zengrui Jin, Mengzhe Geng, Jiajun Deng, Tianzi Wang, Shujie Hu, Guinan Li, and Xunying Liu. 2023a. Personalized adversarial data augmentation for dysarthric and elderly speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Zengrui Jin, Xurong Xie, Mengzhe Geng, Tianzi Wang, Shujie Hu, Jiajun Deng, Guinan Li, and Xunying Liu. 2023b. Adversarial data augmentation using vae-gan for disordered speech recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. Hard negative mixing for contrastive learning. *Advances in neural information processing systems*, 33:21798–21809.

Heejin Kim, Mark Hasegawa Johnson, Jonathan Gunderson, Adrienne Perlman, Thomas Huang, Kenneth Watkin, Simone Frame, Harsh Vardhan Sharma, and Xi Zhou. 2023. Uaspeech.

Yuqin Lin, Longbiao Wang, Jianwu Dang, and Nobuaki Minematsu. 2024. Exploring pre-trained speech model for articulatory feature extraction in dysarthric speech using asr. In *Proceedings of Interspeech 2024*, pages 4598–4602.

Shansong Liu, Xurong Xie, Jianwei Yu, Shoukang Hu, Mengzhe Geng, Rongfeng Su, Shi-Xiong Zhang, Xunying Liu, and Helen Meng. 2020. Exploiting cross-domain visual feature generation for disordered speech recognition. In *Interspeech*, pages 711–715.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Proc. Interspeech 2017*, pages 498–502.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Luke Prananta, Bence Mark Halpern, Siyuan Feng, and Odette Scharenborg. 2022. The effectiveness of time stretching for enhancing dysarthric speech for improved dysarthric speech recognition. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 36–40. ISCA.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Rotem Rousso, Eyal Cohen, Joseph Keshet, and Eleanor Chodroff. 2024. Tradition or innovation: A comparison of modern asr methods for forced alignment. In *Proceedings of Interspeech 2024*, pages 1525–1529, Kos, Greece.

Chetan L Srinidhi and Anne L Martel. 2021. Improving self-supervised learning with hardness-aware dynamic curriculum learning: An application to digital pathology. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 562–571. IEEE.

Helin Wang, Thomas Thebaud, Jesús Villalba, Myra Sydnor, Becky Lammers, Najim Dehak, and Laureano Moro-Velázquez. 2023a. DuTa-VC: A duration-aware typical-to-atypical voice conversion approach with diffusion probabilistic model. In *Interspeech*, pages 1548–1552. ISCA.

Huimeng Wang, Zengrui Jin, Mengzhe Geng, Shujie Hu, Guinan Li, Tianzi Wang, Haoning Xu, and Xunying Liu. 2024a. Enhancing pre-trained asr system fine-tuning for dysarthric speech recognition using adversarial data augmentation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12311–12315. IEEE.

Shiyao Wang, Shiwan Zhao, Jiaming Zhou, Aobo Kong, and Yong Qin. 2024b. Enhancing dysarthric speech recognition for unseen speakers via prototype-based adaptation. In *Proceedings of Interspeech 2024*, pages 1305–1309.

Tianzi Wang, Shoukang Hu, Jiajun Deng, Zengrui Jin, Mengzhe Geng, Yi Wang, Helen Meng, and Xunying Liu. 2023b. Hyper-parameter adaptation of conformer asr systems for elderly and dysarthric speech recognition. In *INTERSPEECH*, pages 1733–1737, Dublin, Ireland. ISCA.

Yiming Wang, Jinyu Li, Heming Wang, Yao Qian, Chengyi Wang, and Yu Wu. 2022. Wav2vec-switch: Contrastive learning from original-noisy speech pairs for robust speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7097–7101. IEEE.

Lidan Wu, Daoming Zong, Shiliang Sun, and Jing Zhao. 2021. A sequential contrastive learning framework for robust dysarthric speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7303–7307. IEEE.

Alex Mihailidis Young. 2010. Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assistive Technology*, 22(2):99–112.

Jianwei Yu, Xurong Xie, Shansong Liu, Shoukang Hu, Max WY Lam, Xixin Wu, Ka Ho Wong, Xunying Liu, and Helen Meng. 2018. Development of the cuhk dysarthric speech recognition system for the ua speech corpus. In *Interspeech*, pages 2938–2942.

Zeyu Zhao and Peter Bell. 2024. Advancing ctc models for better speech alignment: A topological approach. In *IEEE Spoken Language Technology Workshop 2024*, pages 1–7. Institute of Electrical and Electronics Engineers.

Qiu-Shi Zhu, Long Zhou, Jie Zhang, Shu-Jie Liu, Yu-Chen Hu, and Li-Rong Dai. 2023. Robust data2vec: Noise-robust speech representation learning for asr by combining regression and improved contrastive learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

A Appendix

Dysarthria Group	Speaker ID	Speech Intelligibility (%)	# Uttr. (CW/UW)
High (**H**)	F05	95	465/300
	M08	93	465/300
	M10	93	465/300
	M14	90	465/300
	M09	86	465/300
Mid (**M**)	F04	62	**461/289**
	M11	62	465/300
	M05	58	465/300
Low (**L**)	M16	43	465/300
	F02	29	465/300
	M07	28	465/300
Very Low (**VL**)	M01	15	465/300
	M12	7	465/300
	F03	6	**451**/300
	M04	2	465/300

Table 5: Speech intelligibility levels and number of utterances for dysarthric speakers in the UASpeech (Kim et al., 2023) dataset, ordered by intelligibility. The "CW/UW" denotes the number of utterances for common words (CW) and uncommon words (UW). Note that speakers F03 and F04 have fewer utterances.

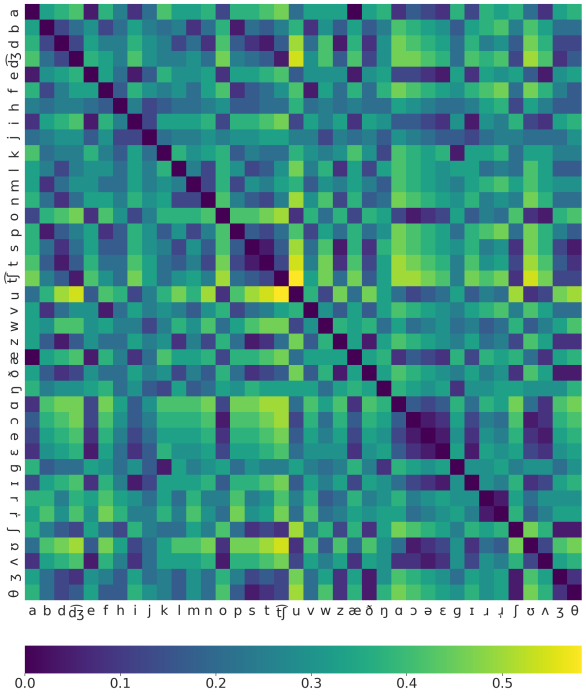

Figure 6: Heat map of phoneme distance matrix (hamming feature edit distance). Brighter areas indicate greater differences in pronunciation between phoneme pair.

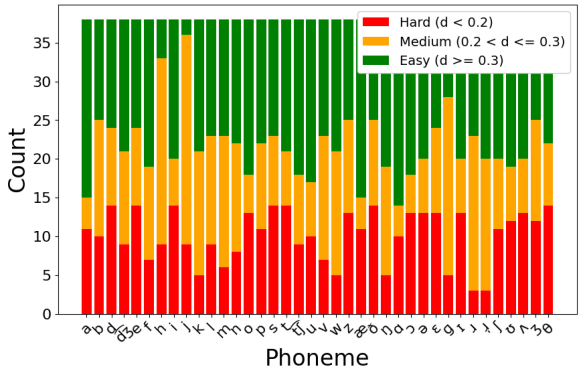

Figure 7: Distribution of difficulty of phoneme pairs in 3 levels: Hard ($d < 0.2$), Medium ($0.2 < d \leq 0.3$), and Easy ($d \geq 0.3$). d is phoneme distance between pairs