

LCS-CTC: Leveraging Soft Alignments to Enhance Phonetic Transcription Robustness

Zongli Ye*, Jiachen Lian[†], Akshaj Gupta[†], Xuanru Zhou*, Haodong Li[‡], Krish Patel[†], Hwi Joo Park[†], Dingkun Zhou[§], Chenxu Guo*, Shuhe Li*, Sam Wang[†], Iris Zhou[†], Cheol Jun Cho[¶], Zoe Ezzes[¶], Jet M.J. Vonk[¶], Brittany T. Morin[¶], Rian Bogley[¶], Lisa Wauters[¶], Zachary A. Miller[¶], Maria Luisa Gorno-Tempini[¶], Gopala Anumanchipalli[†]

*Zhejiang University †UC Berkeley ‡SUSTech §SCUT ¶UCSF

Abstract—Phonetic speech transcription is crucial for fine-grained linguistic analysis and downstream speech applications. While Connectionist Temporal Classification (CTC) is a widely used approach for such tasks due to its efficiency, it often falls short in recognition performance, especially under unclear and nonfluent speech. In this work, we propose LCS-CTC, a two-stage framework for phoneme-level speech recognition that combines a similarity-aware local alignment algorithm with a constrained CTC training objective. By predicting fine-grained frame-phoneme cost matrices and applying a modified Longest Common Subsequence (LCS) algorithm, our method identifies high-confidence alignment zones which are used to constrain the CTC decoding path space, thereby reducing overfitting and improving generalization ability, which enables both robust recognition and text-free forced alignment. Experiments on both LibriSpeech and PPA demonstrate that LCS-CTC consistently outperforms vanilla CTC baselines, suggesting its potential to unify phoneme modeling across fluent and non-fluent speech.

Index Terms—Phoneme, Transcription, Alignment, Clinical

I. INTRODUCTION

Within-word variation in human speech poses challenges for ASR systems [1]–[5] that focus solely on word-level transcription. Yet, modeling subword structures like syllables or phonemes is essential for applications in language learning and clinical analysis [6]–[18]. This work focuses on phoneme transcription—a particularly challenging yet critical task.

Several factors contribute to the difficulty of phoneme-level transcription. First, unlike word-level targets, phoneme annotations are inherently non-deterministic due to accent, allophony and other context-dependent realizations [19]. Second, existing training objectives, such as CTC [20] and attention-based models [21], often introduce alignment noise. For example, CTC tends to produce overly peaked label distributions [22], which result in inaccurate timing and error-prone alignments. Attention-based models, on the other hand, may hallucinate transcriptions [23]—an issue that is particularly problematic in clinical applications where precision is critical. Third, most existing remedies are designed for fluent speech and demonstrate limited effectiveness on non-fluent speech. For example, several approaches [24]–[26] enhance CTC-based alignment by introducing heuristic constraints on the prior.

However, these heuristic alignments have been shown to be incompatible with non-fluent speech [14], [15], [27]–[29].

Given the aforementioned challenges, a key question arises: *Can we design a phoneme transcription objective that performs robustly simultaneously on both fluent and non-fluent speech, while also producing reliable and accurately aligned outputs suitable for clinical applications?*

A key insight arises from treating *phoneme transcription as a speech production process*, particularly when speakers are prompted to read reference text. Fluent speech results from accurate production, while deviations introduce non-fluent speech. For example, when given the phoneme sequence “IH N S ER T,” a speaker might produce “IH [S] N S ER [AH] T,” introducing insertions on “IH” and “ER.” This yields an alignment such as “IH–(IH,S), N–(N), S–(S), ER–(ER,AH), T–(T)” (label–spoken). The core idea is that the *loss function should account only for effective speech-text alignments*. In the case of “IH–(IH,S),” frames aligned to the extraneous “S” should be excluded when computing the loss for “IH.” Without such errors, the method reduces to the standard ASR objective. This alignment strategy follows the Longest Common Subsequence (LCS) principle [30], which was recently shown effective for non-fluent speech [14]. Building on this intuition, we apply the LCS algorithm online to identify effective alignments, which are then used to regularize the vanilla CTC [20] paths, as shown in Fig. 2. We refer to this method as **LCS-CTC**. Unlike previous methods that introduce heuristics to regularize CTC alignments [24]–[26], *LCS-CTC is grounded in human speech production, interpretable, and clinically meaningful*.

Our LCS-CTC framework consists of two main components: a phoneme-aware alignment module based on a modified Longest Common Subsequence (LCS) algorithm, and a constrained CTC training objective that incorporates the resulting alignment masks. The alignment module first predicts a frame-phoneme cost matrix using a lightweight neural network trained with weak supervision. This matrix encodes pairwise acoustic-phonetic similarity and is used to derive a partial alignment path via a similarity-aware LCS dynamic programming procedure. The resulting alignment mask identifies high-confidence frame-phoneme pairs and is used in two complementary ways during training: (1) it anchors specific

frames to phoneme labels with cross-entropy supervision, and (2) it restricts the CTC path space via masked emission probabilities. Together, these two components enable robust phoneme recognition by combining interpretable alignment guidance with flexible sequence modeling.

We evaluate our method on both fluent (LibriSpeech [31]) and non-fluent corpora, including PPA speech [32] and the largest existing simulated non-fluent corpus, *LLM_dys* [33]. Results show that LCS-CTC consistently outperforms vanilla CTC across all metrics, including (weighted) phoneme error rate and duration-aware alignment accuracy. Notably, while LCS-CTC is inspired by non-fluent speech, it also provides stronger regularization for fluent speech. Although we have not yet explored its scalability across data types, linguistic diversity, or domains, the method shows strong potential and may serve as a foundation for universal phoneme recognition. We have open-sourced our model and checkpoints at <https://github.com/Auroraaa86/LCS-CTC>

II. METHODS

Algorithm 1 LCS-based Sequence Alignment

```

Input: Label (phoneme labels),  $C$  (cost matrix),  $tol$  (tolerance),
 $phn\_sim$  (similarity dictionary)
Output: Alignment mask
1: procedure ALIGN(Label,  $C$ ,  $tol$ ,  $phn\_sim$ )
2:    $valid\_res \leftarrow \emptyset$ 
3:   for  $(i, j) \in [0, n] \times [0, m]$  do
4:      $phn1 \leftarrow Label[i]$ ,  $phn2 \leftarrow Label[k]$      $\triangleright k = \arg \min C[:, j]$ 
5:      $sim \leftarrow phn\_sim[(phn1, phn2)]$            $\triangleright$  Lookup similarity
6:     if  $C[i][j] \leq (1 - sim) \times tol$  then       $\triangleright$  Similarity-adjusted
threshold
7:        $valid\_res \leftarrow valid\_res \cup \{(i, j), (j, i)\}$ 
8:     end if
9:   end for
10:   $dp \leftarrow zeros(n + 1, m + 1)$ 
11:  for  $i \leftarrow 1$  to  $n$  do
12:    for  $j \leftarrow 1$  to  $m$  do
13:      if  $(i - 1, j - 1) \in valid\_res$  then
14:         $dp[i][j] \leftarrow \max(dp[i - 1][j - 1] + 1, dp[i][j - 1])$ 
15:        for  $k \leftarrow j + 1$  to  $m$  do           $\triangleright$  Propagate valid matches
16:           $dp[i][k] \leftarrow dp[i][k - 1] + \mathbb{I}_{(i-1, k-1) \in valid\_res}$ 
17:        end for
The indicator function  $\mathbb{I}_{(i-1, k-1) \in valid\_res}$  ensures that alignment
credit is propagated horizontally across frames that are all considered
valid matches to phoneme  $i - 1$ .
18:      else
19:         $dp[i][j] \leftarrow \max(dp[i - 1][j], dp[i][j - 1])$ 
20:      end if
21:    end for
22:  end for
23:   $path \leftarrow \text{TRACEBACK}(dp)$        $\triangleright$  Recover alignment path from DP
24:  return MaskFromPath(path)
25: end procedure

```

The LCS-CTC framework comprises two stages: local alignment extraction via similarity-aware LCS, and constrained CTC training using the resulting alignment masks. This design aims to improve ASR robustness by enforcing phonemically plausible supervision while preserving temporal flexibility.

A. Similarity-Guided LCS Alignment

In this section, we propose a novel method for frame-wise speech-text alignment inspired by the Longest Common

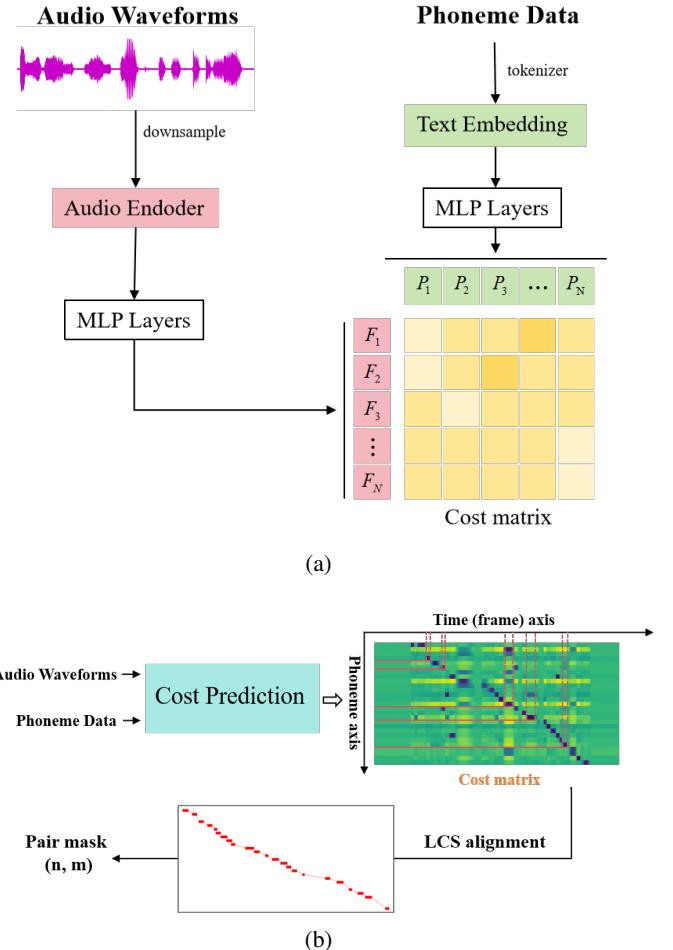


Fig. 1: (a) shows the structure of cost matrix learning model. (b) shows the process from predicted cost matrix to alignment mask: we use cost matrix to construct the dynamic programming matrix and implement modified LCS algorithm to realize a local wise alignment, the dark red dots indicate credible local alignment regions.

Subsequence (LCS) algorithm [30]. Unlike traditional forced alignment approaches that enforce strict one-to-one alignment between each speech frame and a corresponding phoneme label, our method relaxes this assumption by focusing only on “highly confident” frame-label pairs. This partial alignment strategy better reflects the acoustic variability in natural speech, such as ambiguous onsets and offsets, where the phonetic boundaries are inherently fuzzy.

a) Overview: Our method consists of two key stages: (1) predicting a fine-grained cost matrix between speech frames and phoneme labels using a neural model, and (2) applying a similarity-aware dynamic programming alignment inspired by LCS. The goal is to identify optimal frame-label matches that reflect both acoustic evidence and phonetic similarity. The process of our method is shown in Fig. 1

b) Cost Matrix Learning: Given a speech utterance segmented into m frames and its corresponding phoneme sequence of length n , we train a neural model to predict a cost matrix $C \in \mathbb{R}^{n \times m}$ where $C[i, j]$ estimates the alignment

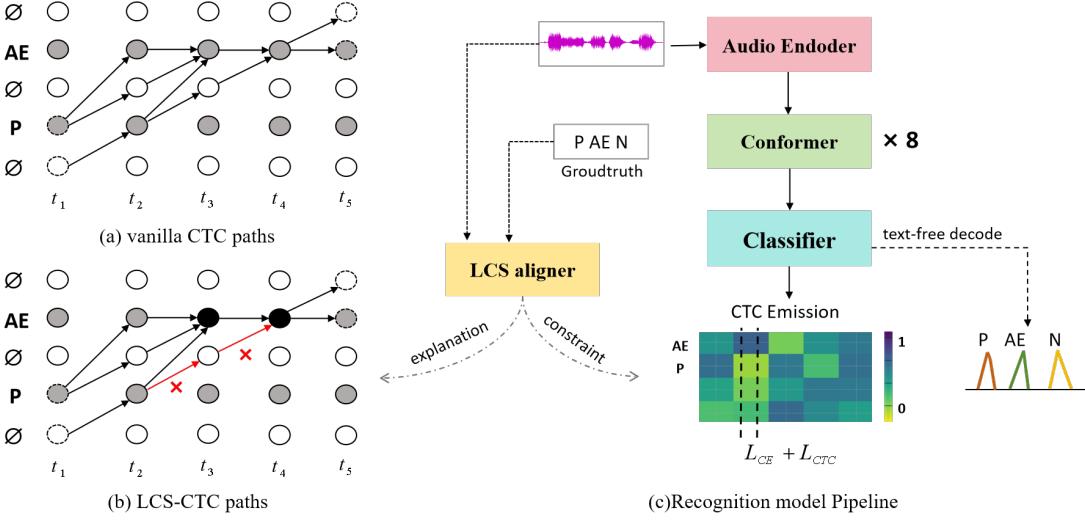


Fig. 2: *Overview of LCS-CTC framework:* (a) Vanilla CTC allows all valid alignment paths that collapse to the target sequence, which can result in peaky distributions. (b) Our LCS-CTC introduces alignment constraints from the LCS-based aligner (black nodes), which eliminate implausible paths (red crosses) by fixing "high-confidence" frame-phoneme matches. (c) The overall training pipeline of LCS-CTC. The LCS aligner uses ground-truth phonemes and predicted emissions to generate alignment masks, which are then used both as supervision (via \mathcal{L}_{CE}) and constraint (via masked \mathcal{L}_{CTC}) during recognition model training.

cost between the i -th phoneme and the j -th frame. The model is trained using a subset of VCTK [34] data, for which accurate frame-level alignments are obtained via Montreal Forced Aligner (MFA) [35] and manually corrected annotations.

To construct target labels for supervision, we incorporate a pre-defined phoneme similarity function $s : \mathcal{P}_{cmu} \times \mathcal{P}_{cmu} \rightarrow [0, 1]$, where \mathcal{P}_{cmu} denotes the set of CMU phonemes. This similarity is computed based on a set of eight articulatory features, including phoneme type (vowel/consonant), vowel length, height, frontness, lip rounding, consonant manner, place of articulation, and voicing. Let the ground-truth aligned phoneme for frame t_j be p_j^* ; then the target cost for matching phoneme p_i to frame t_j is defined as:

$$\tilde{C}[i, j] = \begin{cases} 0, & \text{if } p_i = p_j^*, \\ 1 - s(p_i, p_j^*), & \text{otherwise.} \end{cases}$$

To account for temporal uncertainty at phoneme boundaries (e.g., transitional blur at start or end of a phoneme), we further apply a Gaussian edge attenuation to the ground-truth aligned region. Finally, the cost matrix labels \tilde{C} are normalized across the time axis via softmax. For convenience, we denote $s(p_i, p_j)$ as the similarity values after softmax.

To learn this cost matrix label, we design a model that jointly encodes both the audio signal and the target phoneme sequence. The audio input is first passed through a pre-trained audio encoder (e.g., Wav2Vec2 [36]), producing a sequence of frame-level acoustic embeddings. Simultaneously, the phoneme sequence is mapped into dense embeddings. Each of the two modalities is then processed by separate MLPs to project them into a shared hidden space. The projected frame and phoneme features are combined via negative dot product, followed by a softmax over the time axis. The resulting

matrix is then trained to match the cost matrix labels using a Kullback-Leibler (KL) divergence loss. The structure of the model is illustrated in Fig. 1a.

c) Similarity-Aware LCS Alignment: Once the model yields a predicted cost matrix C , we proceed to perform a similarity-informed dynamic alignment. The alignment algorithm identifies all phoneme-frame pairs (i, j) which the predicted cost $C[i, j]$ falls below a similarity-adjusted threshold $\tau_{i,j} = (1 - s(p_i, p_j)) \cdot \text{tol}$, where tol is a global tolerance hyperparameter, and we set $\text{tol} = 1$ in our work, the selection process is shown in Section III-E. This yields a set of valid match candidates:

$$\text{valid_res} = \{(i, j) \mid C[i, j] \leq (1 - s(p_i, p_k)) \cdot \text{tol}\},$$

$$\text{with } k = \arg \min_i C[i, j].$$

The implementation of the algorithm is shown in Algorithm 1. It outputs binary alignment mask $M' \in \{0, 1\}^{n \times m}$ derived from the alignment path where $M'[i, j] = 1$ if phoneme p_i aligns with frame t_j . This design allows integration into traditional ASR pipelines which we will talk below.

By integrating phonetic similarity into the cost prediction and alignment stages, our method enables more robust alignment in the presence of acoustic ambiguity or phoneme confusability. The process from cost matrix to alignment mask can be seen in Fig. 1b.

B. Constrained LCS-CTC for ASR Training

1) Review on Connectionist Temporal Classification (CTC): The CTC loss has been widely adopted in ASR systems for its ability to model frame-to-label alignments with length mismatch. Given an input sequence of acoustic features $X = \{x_1, x_2, \dots, x_T\}$ and a target label sequence $Y =$

$\{y_1, y_2, \dots, y_L\}$, CTC computes the loss by marginalizing over all possible frame-to-label alignments $\pi \in \mathcal{B}^{-1}(Y)$ where \mathcal{B} is the many-to-one mapping function that removes blank tokens and repeated labels. Vanilla CTC loss is defined as:

$$\mathcal{L}_{\text{CTC}} = -\log \sum_{\pi \in \mathcal{B}^{-1}(Y)} P(\pi | X)$$

While flexible, vanilla CTC suffers from the *peaky distribution* problem, where probability mass tends to concentrate on a few frames, leading to unreliable and non-robust alignments. Moreover, it treats all frame-label alignments as equally likely during training, regardless of their phonetic plausibility, which can harm generalization in downstream recognition tasks.

2) **LCS-CTC: Constrained CTC Training with Alignment Masks:** To enhance the robustness and generalization ability of the recognition model trained by CTC, we propose a novel CTC variant named **LCS-CTC**, which integrates frame-phoneme alignment masks derived from similarity-aware LCS algorithm (as introduced in Section II-A). The central idea is to apply the LCS-derived alignment mask as a constraint over the CTC emission probabilities, thereby pruning unlikely frame-to-phoneme associations during training. In our recognition model, the input audio signal is first encoded by a pretrained speech encoder, followed by a stack of 8 Conformer layers [37] and an MLP-based classifier head that produces the final emission probabilities. Since the emission matrix $P \in \mathbb{R}^{L \times T}$ represents the frame-wise posterior over the entire phoneme vocabulary \mathcal{P}_{cmu} of size L , we transform the original alignment mask M (of shape $n \times T$ for a target sequence of length n) into a binary matrix $M \in \{0, 1\}^{L \times T}$, where $M[i, j] = 1$ indicates that frame t_j is constrained to emit phoneme p_i . This transformation ensures compatibility between the alignment constraint and the CTC output space.

The masked emission \tilde{P} is computed by applying the alignment mask M to the original CTC emission P :

$$\tilde{P}[i, j] = \begin{cases} \frac{P[i, j] \cdot M[i, j] + \epsilon}{\sum_l P[l, j] \cdot M[l, j] + \epsilon}, & \text{if } \sum_l M[l, j] > 0 \\ P[i, j], & \text{otherwise} \end{cases}$$

where ϵ is a small constant to ensure numerical stability. This operation selectively normalizes emission probabilities at frames that are constrained by the LCS-derived alignment mask, while leaving unconstrained frames unchanged.

We then define a hybrid training objective that supervises the model in two complementary ways. First, for the high-confidence alignment regions specified by the binary mask M , we apply cross-entropy loss directly on the original emission P to provide strong local supervision. Second, we apply vanilla CTC loss over the full sequence using the masked emission \tilde{P} , which softly restricts the CTC decoding space during sequence-level training. Crucially, by anchoring certain frames to fixed phoneme labels through M , our method effectively eliminates implausible alignment paths and narrows the search space, thereby improving CTC training stability and generalization. The details are illustrated in Fig. 2.

The final loss is computed as:

$$\mathcal{L}_{\text{LCS-CTC}} = \lambda \cdot \mathcal{L}_{\text{CE}}(P \odot M, Y) + (1 - \lambda) \cdot \mathcal{L}_{\text{CTC}}(\tilde{P}, Y)$$

where \odot denotes element-wise masking. Here, \mathcal{L}_{CE} denotes standard cross-entropy loss evaluated only on masked positions, while $\lambda \in [0, 1]$ is a weighting factor balancing between strict alignment enforcement and flexible CTC decoding. We set $\lambda = 0.5$ in our work. This hybrid objective ensures that the model is supervised on phonemically plausible alignments while retaining the benefits of sequence-level flexibility.

III. EXPERIMENTS

A. Datasets

We use **VCTK** [34] dataset which includes 109 native English speakers with accented speech to train our cost matrix learning model and phoneme recognition model, and we conduct experiments on three ASR datasets to evaluate our proposed LCS-CTC’s performance: (1) **Librispeech** [31], which contains 1000 hours of English speech, we use its *dev* and *test* subsets for evaluation. (2) **PPA Speech** [32], it is collected in collaboration with clinical experts and includes recordings from 38 participants diagnosed with Primary Progressive Aphasia (PPA). Participants were asked to read the “grandfather passage,” resulting in approximately one hour of speech in total. (3) **LLM_dys**, as mentioned in [33], We used a large language model (claude-3-5-sonnet-20241022 [38]) to create a synthetic speech dataset. Given clean text and its CMU/IPA sequences, the model was prompted to insert dysfluencies naturally. This produced dysfluent IPA sequences for VITS-based speech synthesis [39], as well as CMU sequences labeled with dysfluencies for evaluation.

B. Implementation details

In our main experiments, we split the VCTK dataset into four partitions with a ratio of 3:1:5:1. Specifically, 30% of the data is used for training the cost matrix learning model. These samples are annotated with MFA and manual refinement, mentioned in Section II-A. An additional 10% is held out as the validation set for this stage.

For training the phoneme recognition model based on our proposed LCS-CTC, we use 50% of the data for training and 10% for validation. All experiments are conducted on an NVIDIA A6000 GPU. We use a learning rate of $1e-5$ and a batch size of 1 for all training stages. No weight decay is applied during optimization.

C. Evaluation Metrics

1) **Phoneme Error Rate (PER):** Phoneme Error Rate is a common metric for evaluating phoneme-level recognition performance. It is defined as the Levenshtein distance between the predicted and reference phoneme sequences, normalized by the length of the reference. The metric accounts for substitution, insertion, and deletion errors, and reflects the overall transcription accuracy at the phonetic level.

Model	Method	LibriSpeech				PPA			LLM_dys
		test-clean	test-other	dev-clean	dev-other	nfvppa	lvppa	svppa	
Wav2Vec2.0-L	CTC	16.16	21.04	15.41	20.50	41.04	36.63	15.78	14.52
	CE-CTC	16.84	21.16	15.73	20.68	38.89	38.75	17.18	14.75
	LCS-CTC(ours)	16.09	20.94	15.10	19.70	38.14	32.04	15.01	13.89
HuBERT-L	CTC	15.00	18.16	14.12	17.18	32.70	18.34	13.32	12.13
	CE-CTC	14.83	17.75	14.02	17.07	34.45	19.06	15.24	12.94
	LCS-CTC(ours)	14.75	17.35	11.78	16.72	32.02	17.71	13.30	11.85
WavLM-L	CTC	12.62	15.01	13.92	14.99	29.88	20.33	14.14	12.13
	CE-CTC	12.31	14.73	12.83	14.90	30.34	22.15	15.05	11.30
	LCS-CTC(ours)	12.09	14.45	12.41	14.73	29.51	18.72	13.83	11.22

TABLE I: PER(%) performance of three methods on LibriSpeech, PPA, and LLM_dys dataset.

Model	Method	Librispeech				PPA			LLM_dys
		test-clean	test-other	dev-clean	dev-other	nfvppa	lvppa	svppa	
WavLM-L	CTC	10.75	13.41	11.96	12.07	23.14	11.46	11.70	6.26
	CE-CTC	10.73	13.41	11.70	12.14	23.55	13.05	12.94	5.71
	LCS-CTC(ours)	10.49	12.85	11.56	11.97	22.75	10.92	11.00	5.64

TABLE II: WPER(%) performance of three methods on LibriSpeech, PPA, and LLM_dys dataset.

Model	Method	Librispeech				LLM_dys
		test-clean	test-other	dev-clean	dev-other	
WavLM-L	CTC	146.6	113.5	115.8	131.3	95.1
	LCS-CTC(ours)	115.8	94.5	94.9	102.9	76.7

TABLE III: BL(ms) performance of different methods on LibriSpeech and LLM_dys dataset.

Model	Method	Librispeech				LLM_dys
		test-clean	test-other	dev-clean	dev-other	
WavLM-L	CTC	22.4	14.8	19.7	28.0	17.9
	LCS-CTC(ours)	16.0	10.8	13.1	21.9	12.7

TABLE IV: ARL performance of different methods on LibriSpeech and LLM_dys dataset.

2) *Weighted Phonetic Error Rate (WPER)*: Unlike PER which assumes all phonemes are equally distant, WPER takes phonetic variation into account by incorporating phoneme similarity. This addresses the shortcomings of traditional error metrics and offers a more accurate evaluation of recognition performance. Refer to [40], we define WPER as

$$\text{WPER} = \frac{\sum_{(p_t, p_s)} (1 - s(p_t, p_s)) + D + I}{N}$$

where p_t, p_s denote target and substitute phonemes, s is their similarity score defined in II-A, and N is the reference phoneme sequence length. D and I represent deletion and insertion errors respectively. For substitutions, phoneme similarity is evaluated first, followed by error penalties.

3) *Boundary Loss (BL)*: Boundary Loss (BL) quantifies the alignment quality between predicted and ground-truth phoneme boundaries. Specifically, it is computed as the average absolute deviation between the predicted and reference onset/offset times of each phoneme. A lower BL reflects more accurate temporal alignment.

4) *Articulatory Reconstruction Loss (ARL)*: To assess phoneme-level articulatory consistency, we introduce ARL. It measures the L2 distance between articulatory embeddings reconstructed from predicted phonemes and those extracted from the original audio using an acoustic-to-articulatory inversion

(AAI) model [41]. Lower ARL indicates better articulatory plausibility of the recognition output.

D. Results and Discussion

Table I presents the PER of different training methods, including vanilla CTC and our proposed LCS-CTC—applied to three common speech encoders: Wav2Vec2.0 [36], HuBERT [42], and WavLM [43]. Across all models and datasets, LCS-CTC outperforms vanilla CTC on both clean (e.g., test-clean, dev-clean) and challenging conditions (e.g., test-other, lvppa, LLM_dys), demonstrating strong generalization.

We also compare with CE-CTC, a variant that combines CTC with frame-level cross-entropy loss derived from ground-truth alignments. While CE-CTC introduces additional frame-level supervision, it does not consistently outperform vanilla CTC, especially on more challenging datasets such as PPA and LLM_dys. This suggests that relying on full-frame labels does not always lead to better generalization, and may even introduce noise under dysfluent or noisy speech conditions. In contrast, LCS-CTC achieves competitive or superior performance without relying on frame-level labels at training time, making it more practical and scalable.

Furthermore, we report WPER in Table II using the best-performing encoder WavLM. LCS-CTC again shows the best performance on all subsets. These results confirm that our

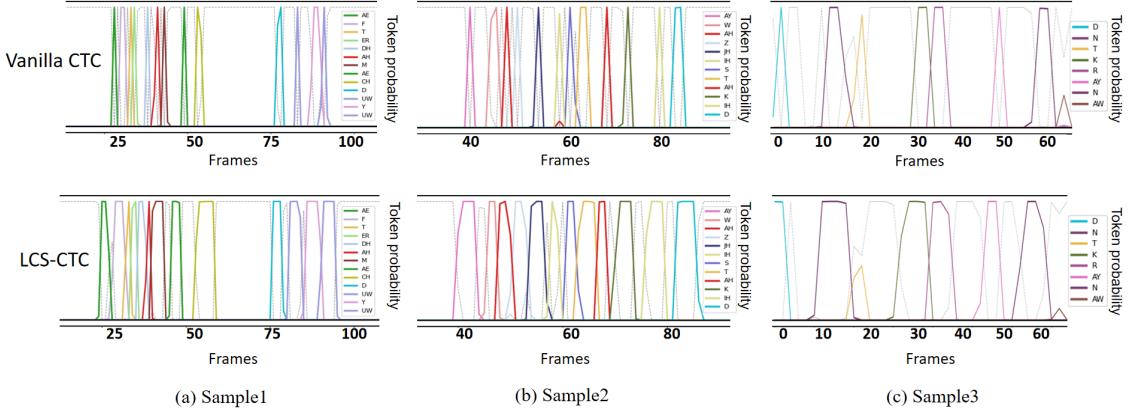


Fig. 3: Visualization of token emission probabilities for vanilla CTC and our proposed LCS-CTC on three randomly selected samples from the LibriSpeech and PPA. The gray dashed lines indicate the blank token. Compared to vanilla CTC, LCS-CTC produces smoother and more distributed token activations, characterized by more consistent repetitions of non-blank tokens.

method leads to more accurate and robust phoneme recognition across various speech conditions.

We use Boundary Loss (BL) to measure alignment quality. As shown in Table III, LCS-CTC significantly reduces BL compared to vanilla CTC, indicating more accurate and temporally consistent alignments. This is further supported by the emission visualizations in Fig. 3, where LCS-CTC produces smoother and less peaky token distributions. Such smoother emission behavior enhances generalization and likely contributes to the improved recognition performance observed earlier. Moreover, our model can also serve as a text-free phoneme-level aligner. The results in Table IV demonstrate that LCS-CTC yields consistently lower articulatory reconstruction loss compared to vanilla CTC. This suggests that the phoneme sequences predicted by LCS-CTC better preserve articulatory smoothness and continuity, leading to embeddings that more closely match the natural trajectories derived from the speech signal.

E. Ablation experiments

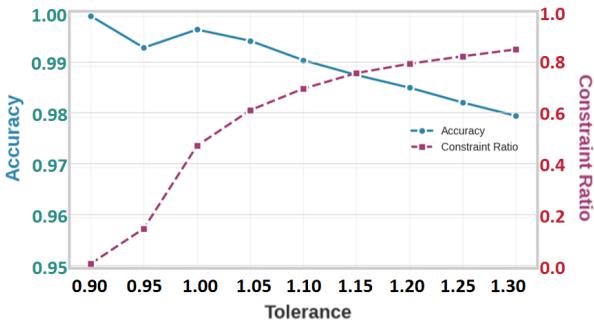


Fig. 4: Effect of tol on alignment accuracy and constrained region ratio.

We perform an ablation study on the tolerance parameter tol in the LCS alignment algorithm, which controls the threshold for accepting frame-phoneme matches based on predicted cost and phoneme similarity. Smaller values of tol impose

stricter alignment conditions, leading to higher alignment precision but fewer constrained regions. This may reduce the effectiveness of the frame-level supervision in our LCS-CTC. In contrast, larger tol values include more frames but risk introducing noisy or imprecise alignments, which can weaken training. We evaluate values in the range $\{0.9 - 1.3\}$ and observe that $tol = 1.0$ provides the best balance between alignment accuracy and coverage, and is thus adopted in all main experiments.

IV. CONCLUSIONS AND LIMITATIONS

In this work, we propose LCS-CTC which demonstrates strong performance across all evaluation metrics for both fluent and non-fluent speech, and, critically, yields outputs that are more clinically interpretable. Nonetheless, several limitations remain. The current framework is trained on clean speech from the VCTK corpus, which is insufficient to fully assess its generalizability to more diverse, noisy, or pathological speech. Future work will focus on expanding both the training and evaluation datasets to encompass a broader range of conditions. LCS-CTC operates explicitly as a forced aligner, maintaining transparency comparable to traditional HMM-based systems such as MFA [35]. A promising direction for future research is the development of a neural variant [3], [44]–[46] that retains interpretability while achieving performance on par with established HMM-based methods. Another challenge involves the inherent ambiguity of phonetic labels. To address this, we plan to incorporate phonetic similarity modeling [47] and WFST-based decoding [40], [48] strategies. Moreover, given that phonemes are intrinsically articulatory in nature, we aim to integrate articulatory priors [14], [49]–[51] into the label space to mitigate alignment instability—particularly in disordered or atypical speech.

V. ACKNOWLEDGEMENTS

Thanks for support from UC Noyce Initiative, Society of Hellman Fellows, NIH/NIDCD, and the Schwab Innovation fund. Many thanks to Alexei Baevski for the early-stage discussions and insightful idea brainstorming.

REFERENCES

- [1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 28492–28518.
- [2] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang *et al.*, “Google usm: Scaling automatic speech recognition beyond 100 languages,” *arXiv preprint arXiv:2303.01037*, 2023.
- [3] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi *et al.*, “Scaling speech technology to 1,000+ languages,” *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.
- [4] K. C. Puvvada, P. Źelasko, H. Huang, O. Hrinchuk, N. R. Koluguri, K. Dhawan, S. Majumdar, E. Rastorgueva, Z. Chen, V. Lavrukhin *et al.*, “Less is more: Accurate speech recognition & translation without web-scale data,” *arXiv preprint arXiv:2406.19674*, 2024.
- [5] NVIDIA NeMo and Suno.ai, “Parakeet-rnnt-1.1b: A fastconformer transducer model for english asr,” <https://huggingface.co/nvidia/parakeet-rnnt-1.1b>, 2025.
- [6] Y. El Kheir, A. Ali, and S. A. Chowdhury, “Automatic pronunciation assessment - a review,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8304–8324.
- [7] K. Truong, A. Neri, C. Cucchiari, and H. Strik, “Automatic pronunciation error detection: an acoustic-phonetic approach,” in *InSTIL/ICALL 2004 Symposium on Computer Assisted Learning*, 2004, p. paper 032.
- [8] A. Lee *et al.*, “Language-independent methods for computer-assisted pronunciation training,” Ph.D. dissertation, Massachusetts Institute of Technology, 2016.
- [9] C. Cordella, L. Di Filippo, V. B. Kolachalam, and S. Kiran, “Connected speech fluency in poststroke and progressive aphasia: A scoping review of quantitative approaches and features,” *American Journal of Speech-Language Pathology*, vol. 33, no. 4, pp. 2083–2120, 2024.
- [10] L. Fontan, T. Prince, A. Nowakowska, H. Sahraoui, and S. Martinez-Ferreiro, “Automatically measuring speech fluency in people with aphasia: First achievements using read-speech data,” *Aphasiology*, vol. 38, no. 5, pp. 939–956, 2024.
- [11] J. K. Gordon, “The fluency dimension in aphasia,” *Aphasiology*, vol. 12, no. 7–8, pp. 673–688, 1998.
- [12] J. K. Gordon and S. Clough, “How do clinicians judge fluency in aphasia?” *Journal of Speech, Language, and Hearing Research*, vol. 65, no. 4, pp. 1521–1542, 2022.
- [13] J. Metu, V. Kotha, and A. E. Hillis, “Evaluating fluency in aphasia: Fluency scales, trichotomous judgements, or machine learning,” *Aphasiology*, pp. 1–13, 2023.
- [14] J. Lian, X. Zhou, Z. Ezzes, J. Vonk, B. Morin, D. P. Baquirin, Z. Miller, M. L. Gorno Tempini, and G. Anumanchipalli, “Ssdm: Scalable speech dysfluency modeling,” in *Advances in Neural Information Processing Systems*, vol. 37, 2024.
- [15] J. Lian, X. Zhou, C. Guo, Z. Ye, Z. Ezzes, J. Vonk, B. Morin, D. Baquirin, Z. Mille, M. L. G. Tempini, and G. K. Anumanchipalli, “Automatic detection of articulatory-based disfluencies in primary progressive aphasia,” *IEEE JSTSP*, 2025.
- [16] X. Zhou, A. Kashyap, S. Li, A. Sharma, B. Morin, D. Baquirin, J. Vonk, Z. Ezzes, Z. Miller, M. Tempini, J. Lian, and G. Anumanchipalli, “Yolo-stutter: End-to-end region-wise speech dysfluency detection,” in *Interspeech 2024*, 2024, pp. 937–941.
- [17] X. Zhou, C. J. Cho, A. Sharma, B. Morin, D. Baquirin, J. Vonk, Z. Ezzes, Z. Miller, B. L. Tee, M. L. Gorno-Tempini *et al.*, “Stutter-solver: End-to-end multi-lingual dysfluency detection,” in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 1039–1046.
- [18] X. Zhou, J. Lian, C. J. Cho, J. Liu, Z. Ye, J. Zhang, B. Morin, D. Baquirin, J. Vonk, Z. Ezzes, Z. Miller, M. L. G. Tempini, and G. Anumanchipalli, “Time and tokens: Benchmarking end-to-end speech dysfluency detection,” 2024. [Online]. Available: <https://arxiv.org/abs/2409.13582>
- [19] K. Choi, E. Yeo, K. Chang, S. Watanabe, and D. Mortensen, “Leveraging allophony in self-supervised speech models for atypical pronunciation assessment,” in *NAACL*, 2025.
- [20] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.
- [21] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” *Advances in neural information processing systems*, vol. 28, 2015.
- [22] A. Zeyer, R. Schlüter, and H. Ney, “Why does ctc result in peaky behavior?” *arXiv preprint arXiv:2105.14849*, 2021.
- [23] A. Koenecke, A. S. G. Choi, K. X. Mei, H. Schellmann, and M. Sloane, “Careless whisper: Speech-to-text hallucination harms,” in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024, pp. 1672–1681.
- [24] J. Tian, B. Yan, J. Yu, C. Weng, D. Yu, and S. Watanabe, “Bayes risk ctc: Controllable ctc alignment in sequence-to-sequence tasks,” *ICLR*, 2023.
- [25] R. Huang, X. Zhang, Z. Ni, L. Sun, M. Hira, J. Hwang, V. Manohar, V. Pratap, M. Wiesner, S. Watanabe *et al.*, “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)*. IEEE, 2024, pp. 11831–11835.
- [26] Z. Yao, W. Kang, X. Yang, F. Kuang, L. Guo, H. Zhu, Z. Jin, Z. Li, L. Lin, and D. Povey, “Cr-ctc: Consistency regularization on ctc for improved speech recognition,” *ICLR*, 2025.
- [27] J. Lian, C. Feng, N. Farooqi, S. Li, A. Kashyap, C. J. Cho, P. Wu, R. Netzorg, T. Li, and G. K. Anumanchipalli, “Unconstrained dysfluency modeling for dysfluent speech transcription and detection,” in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.
- [28] J. Lian and G. Anumanchipalli, “Towards hierarchical spoken language disfluency modeling,” in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.
- [29] Z. Ye, J. Lian, X. Zhou, J. Zhang, H. Li, S. Li, C. Guo, A. Das, P. Park, Z. Ezzes, J. Vonk, B. Morin, R. Bogley, L. Wauters, Z. Miller, M. Gorno-Tempini, and G. Anumanchipalli, “Seamless dysfluent speech text alignment for disordered speech analysis,” *Interspeech*, 2025.
- [30] D. S. Hirschberg, “Algorithms for the longest common subsequence problem,” *Journal of the ACM (JACM)*, vol. 24, no. 4, pp. 664–675, 1977.
- [31] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [32] M. L. Gorno-Tempini, A. E. Hillis, S. Weintraub, A. Kertesz, M. Mendez, S. F. Cappa, J. M. Ogar, J. D. Rohrer, S. Black, B. F. Boeve *et al.*, “Classification of primary progressive aphasia and its variants,” *Neurology*, vol. 76, no. 11, pp. 1006–1014, 2011.
- [33] J. Zhang, X. Zhou, J. Lian, S. Li, W. Li, Z. Ezzes, R. Bogley, L. Wauters, Z. Miller, J. Vonk, B. Morin, M. Gorno-Tempini, and G. Anumanchipalli, “Analysis and evaluation of synthetic data generation in speech dysfluency detection,” *Interspeech*, 2025.
- [34] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, “Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92),” *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2019.
- [35] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal forced aligner: Trainable text-speech alignment using kaldii,” in *Interspeech 2017*, 2017, pp. 498–502.
- [36] 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*, 2020.
- [37] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, “Conformer: Convolution-augmented transformer for speech recognition,” *Interspeech*, 2020.
- [38] Anthropic, “Model card addendum: Claude 3.5 haiku and upgraded claude 3.5 sonnet,” 2024. [Online]. Available: <https://www.anthropic.com>
- [39] J. Kim, J. Kong, and J. Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.
- [40] C. Guo, J. Lian, X. Zhou, J. Zhang, S. Li, Z. Ye, H. J. Park, A. Das, Z. Ezzes, J. Vonk, B. Morin, R. Bogley, L. Wauters, Z. Miller, M. Gorno-Tempini, and G. Anumanchipalli, “Dysfluent wfst: A framework for zero-shot speech dysfluency transcription and detection,” *Interspeech*, 2025.
- [41] C. J. Cho, P. Wu, T. S. Prabhune, D. Agarwal, and G. K. Anumanchipalli, “Coding speech through vocal tract kinematics,” *IEEE*

- Journal of Selected Topics in Signal Processing*, vol. 18, no. 8, p. 1427–1440, Dec. 2024. [Online]. Available: <http://dx.doi.org/10.1109/JSTSP.2024.3497655>
- [42] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE TASLP*, 2021.
 - [43] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE JSTSP*, 2022.
 - [44] J. Zhu, C. Zhang, and D. Jurgens, “Phone-to-audio alignment without text: A semi-supervised approach,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8167–8171.
 - [45] J. Li, Y. Meng, Z. Wu, H. Meng, Q. Tian, Y. Wang, and Y. Wang, “Neufa: Neural network based end-to-end forced alignment with bidirectional attention mechanism,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8007–8011.
 - [46] “Joint speech and text machine translation for up to 100 languages,” *Nature*, vol. 637, no. 8046, pp. 587–593, 2025.
 - [47] X. Zhou, J. Lian, C. J. Cho, T. Prabhune, S. Li, W. Li, R. Ortiz, Z. Ezzes, J. Vonk, B. Morin, R. Bogley, L. Wauters, Z. Miller, M. Gorno-Tempini, and G. Anumanchipalli, “Towards accurate phonetic error detection through phoneme similarity modeling,” *Interspeech*, 2025.
 - [48] S. Li, C. Guo, J. Lian, C. J. Cho, W. Zhao, X. Zhou, D. Zhou, S. Wang, G. Wang, J. Yang *et al.*, “K-function: Joint pronunciation transcription and feedback for evaluating kids language function,” *arXiv preprint arXiv:2507.03043*, 2025.
 - [49] C. J. Cho, P. Wu, T. S. Prabhune, D. Agarwal, and G. K. Anumanchipalli, “Coding speech through vocal tract kinematics,” in *IEEE JSTSP*, 2025.
 - [50] J. Lian, A. W. Black, L. Goldstein, and G. K. Anumanchipalli, “Deep Neural Convulsive Matrix Factorization for Articulatory Representation Decomposition,” in *Proc. Interspeech 2022*, 2022, pp. 4686–4690.
 - [51] J. Lian, A. W. Black, Y. Lu, L. Goldstein, S. Watanabe, and G. K. Anumanchipalli, “Articulatory representation learning via joint factor analysis and neural matrix factorization,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.