Subphonetic Acoustic Modeling via Optimal Transport for Pronunciation Assessment

Double Blind Review

***Abstract*—Pronunciation assessment requires acoustic evidence that is temporally precise, diagnostically meaningful, and faithful to the learner's actual production. However, existing acoustic models often struggle to provide recognition and segmentation evidence simultaneously. CTC-based phone recognizers can predict phone sequences flexibly, but their sparse and peaky posteriors often miss phone boundaries and fine-grained pronunciation cues. In contrast, text-dependent forced aligners provide reliable temporal information when transcripts are available, but are not directly applicable to reference-free pronunciation analysis. In this work, we propose a topology-aware frame-wise acoustic model that learns dense ordered state posteriors within each phone. The key idea is to recover phone-internal state structure in a neural acoustic model by combining ordered subphonetic states with optimal temporal transport classification (OTTC). This combination encourages dense monotonic frame-level state discrimination while preserving phone recognition ability. Experiments on read, spontaneous, and L2 speech show improved segmentation over neural baselines with competitive recognition performance. Downstream evaluations further show gains in mispronunciation detection and automatic pronunciation assessment. Probing analysis suggests that the learned states capture phoneme-dependent acoustic structure rather than arbitrary frame-level distributions.**[1]

***Index Terms*—pronunciation assessment, forced alignment, phone segmentation, acoustic modeling, connectionist temporal classification, optimal transport**

I. Introduction

Pronunciation assessment (PA) aims to provide learners with fine-grained feedback on how their speech deviates from target-language production [1]. This requires an acoustic model to answer not only which phone was produced, but also where it was realized and how its internal acoustic trajectory departs from the target. Such temporally precise and phone-internal evidence supports goodness-of-pronunciation (GOP) scoring [2], mispronunciation detection and diagnosis (MDD) [3], [4], and automatic pronunciation assessment (APA) [5], [6], which are crucial for computer aided pronunciation training (CAPT) tasks.

This requirement exposes a tension between recognition and segmentation. Forced alignment (FA) models can produce reliable phone boundaries when a transcript is provided [7]–[9], but they do not provide reliable phone hypotheses in reference-free settings. In contrast, neural phone recognition (PR) models can predict phone sequences without a fixed transcript, but often discard fine-grained temporal dynamics and phone boundaries.

Connectionist temporal classification (CTC) based PR is a representative example of this trade-off [10]. By marginalizing over possible frame-to-label alignments, CTC enables phone-sequence prediction without explicit frame-level supervision [10]. However, although effective under general PR tasks, CTC has been reported to be suboptimal in pronunciation-assessment-related tasks such as MDD [11], [12]. This is because subphonetic cues, which often reveal non-canonical pronunciation patterns, are largely neglected by the CTC criterion. Moreover, the blank-dominated and peaky posteriors of CTC [13]–[15] make it difficult to predict precise phone boundaries as well.

To address these problems, we introduce a topology-aware frame-wise acoustic model trained with optimal temporal transport classification (OTTC) [16]. The model expands each phone into ordered subphonetic states and learns dense monotonic frame-to-state assignments, bringing HMM-style phone-internal topology into neural pronunciation modeling. Rather than optimizing PR accuracy alone, the model is also designed to provide temporally dense acoustic evidence for PA. The main contributions are as follows:

- We propose a topology-aware frame-wise acoustic model that introduces temporally ordered subphonetic states for PA.
- We combine the proposed topology with OTTC to obtain dense monotonic alignments while retaining competitive PR performance.
- We evaluate the model on read speech, spontaneous speech, and L2 speech, covering both text-dependent forced alignment (TDFA) and text-independent (TI) segmentation/recognition.
- We further validate the learned acoustic evidence on L2 PA tasks, and analyze how the learned topology states represent phoneme-dependent characteristics.

II. Research Background

In this section, we revisit conventional forced alignment methods and neural phonetic aligners, and then explain the motivation behind the proposed framework.

A. Recognition–segmentation Trade-off

As discussed in Sec. I, current acoustic modeling approaches often struggle to provide precise temporal localization and flexible phone recognition at the same time. For example, classical forced alignment systems based on hidden Markov models (HMMs) recover a dense monotonic frame-to-phone path from a given transcript and can provide stable phone boundaries [7]–[9]. Their strength comes from using the reference sequence as a strong constraint, where each frame

[1]Code and checkpoints will be released upon acceptance.

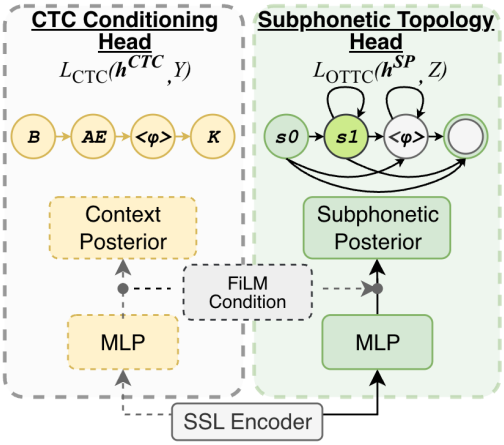


Fig. 1: Illustration of proposed subphonetic model.

is assigned to a corresponding HMM state. This constraint supports accurate boundary estimation, but also makes decoding less flexible in reference-free settings. In contrast, CTC-based phone recognizers can infer phone sequences under weaker constraints, but their repeat-and-collapse objective encourages sparse non-blank emissions, making them unreliable for boundary estimation [13]. Several CTC variants have attempted to mitigate this limitation. For example, Huang et al. introduced label priors to dynamically suppress blank probabilities [17]. However, their improvements in segmentation accuracy remain limited.

Recent neural-network-based alignment models attempt to move beyond this trade-off by recovering better alignment behavior while retaining recognition ability. Charsiu [18] is a representative example: it uses self-supervised speech representations (SSL), forward-sum loss [19], curriculum learning, and frame-wise cross-entropy supervision to improve phone-to-audio alignment [18]. This line of work shows that neural acoustic models can balance recognition and segmentation more effectively. Meanwhile, as neural networks already provide better phone-level recognition performance, less attention has been paid to modeling acoustic structure below the phoneme level.

B. Subphonetic Acoustic Modeling

Another reason HMM-based models remain effective for segmentation is their use of multiple ordered states for each acoustic unit. In HMM-based ASR, a phone is commonly represented by several left-to-right (LTR) Markov states, often using a three-state topology. This structure allows the model to represent coarse phone-internal structures such as onset, steady-state, and offset regions [7], [8], [20]. Fenone [21] and Senone [22] further refined this idea by clustering context-dependent subphonetic states for recognition.

From this perspective, standard CTC can be viewed as a much simpler HMM, where each phone is modeled as a single label state together with the blank token. Recent CTC-topology studies also restore richer state graphs to suppress the blank ratio and improve alignment behavior [23], [24]. Thus, introducing subphonetic states is not a new concept. Rather, we

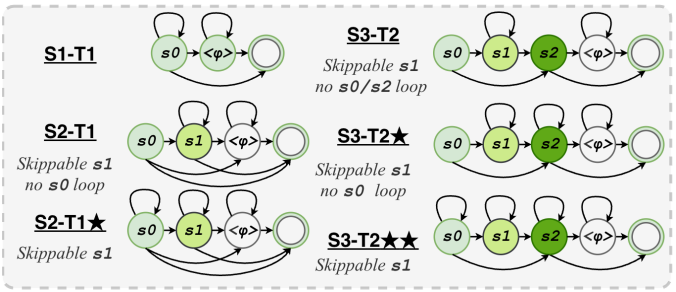


Fig. 2: Illustration of subphonetic topologies.

revisit ordered subphonetic topology as a way to provide dense and diagnostic acoustic evidence for PA. To better exploit the advantages of subphonetic modeling, we introduce a topology-aware OTTC formulation in the next section.

III. Topology-Aware Subphonetic Modeling via Optimal Temporal Transport Classification

In this section, we explain standard CTC and extend it to ordered multi-state subphonetic topologies. We then analyze why topology alone remains insufficient under a sparse CTC criterion, and introduce OTTC as monotonic frame-level training for learning ordered subphonetic states.

A. CTC and Subphonetic Topology

Given an acoustic frame sequence $\mathbf{X} = \{x_i\}_{i=1}^{n}$ and a monophone target sequence $\mathbf{Y} = (y_1, \ldots, y_m)$, CTC optimizes the marginal probability over all possible frame-level paths that collapse to $\mathbf{Y}$:

$$\mathcal{L}_{\text{CTC}} = -\log \sum_{\pi \in \mathcal{B}^{-1}(\mathbf{Y})} \prod_{i=1}^{n} p(\pi_i \mid x_i), \tag{1}$$

where $\mathcal{B}^{-1}(\mathbf{Y})$ denotes the set of all paths that collapse to $\mathbf{Y}$.

Recent work has also explored CTC-based architectures by changing the topology of the output units [23], [24]. Following the topology notation in [23] and the illustration in Fig. 2, we denote the topology of an acoustic unit as SxTy. Sx denotes the number of ordered states used to represent the unit, while Ty specifies the minimum number of states that any valid traversal must consume. The $\star$ symbols add self-loops to selected states, allowing those states to occupy additional consecutive frames without changing the minimum traversal length. In this view, S1T1 corresponds to standard one-state CTC phone modeling, and S3T2$\star\star$ means that each phone can be represented by three ordered states with a minimum two-state traversal during decoding. Let $\mathcal{P}$ denote the phone inventory. For each phone $y \in \mathcal{P}$, a topology expansion function maps it into ordered subphonetic states:

$$\tau(y) = \left(y^{(1)}, y^{(2)}, \ldots, y^{(K_y)}\right), \tag{2}$$

where K_y is determined by the chosen topology. The monophone sequence $\mathbf{Y}$ is then expanded into a subphonetic state sequence:

$$\mathbf{Z} = \tau(\mathbf{Y}) = \tau(y_1) \oplus \tau(y_2) \oplus \cdots \oplus \tau(y_m), \tag{3}$$

where $\oplus$ denotes sequence concatenation.

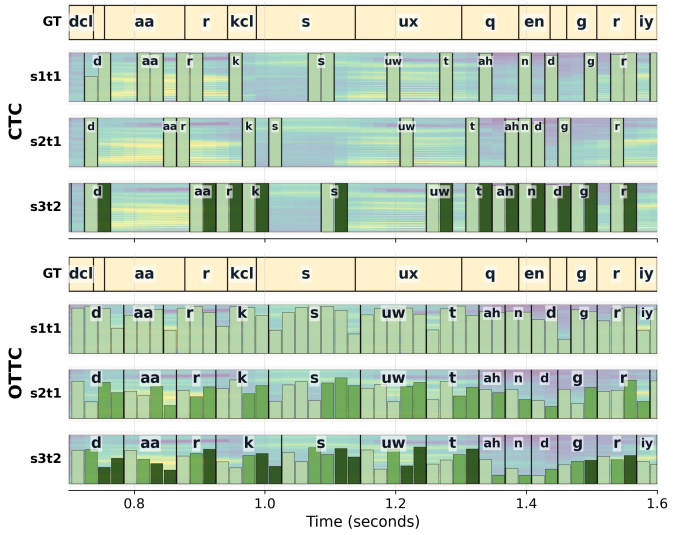


Fig. 3: Comparison of frame-level probability distributions across training criteria and topologies. Different topology states are shown in different colors. CTC produces sparse spikes, while OTTC produces a dense and complete subphonetic trace.

B. Limitations of Topology under CTC Training

Although topology states provide a richer label space, they do not by themselves guarantee reliable state transitions. Multi-state CTC topologies can reduce the blank ratio, but neighboring topology states may still collapse in practice [24]. As illustrated in Fig. 3, even when a phone is expanded into S3 states, its posterior may cover only part of the ordered state sequence. Thus, although topology specifies the possible LTR state order, the CTC objective can still satisfy sequence prediction without assigning dense evidence to every state. This makes the resulting posteriors still unreliable for boundary estimation and fine-grained pronunciation analysis.

C. OTTC as Dense Monotonic State Transport

To address this limitation, we introduce OTTC [16]. OTTC formulates alignment as a one-dimensional mass transport problem. Instead of marginalizing over sparse CTC paths, it transports frame-level source mass to the expanded topology-state sequence $\mathbf{Z}$ and optimizes the posterior under this dense monotonic transport plan. This encourages the ordered states to receive frame-wise acoustic evidence rather than only isolated recognition peaks. Let $\mathbf{Z} = (z_1, \ldots, z_M)$ denote the expanded subphonetic state sequence derived from $\mathbf{Y}$. The frame weights $\alpha \in \Delta^n$ are predicted from the acoustic sequence by a neural network W:

$$\alpha[\mathbf{X}, W] = \mathrm{softmax}(W(x_1), \ldots, W(x_n))^\top, \tag{4}$$

and the target weights $\beta \in \Delta^M$ are typically fixed as a uniform distribution over target labels. The alignment plan $\gamma(\alpha, \beta)$ is obtained by solving

$$\gamma(\alpha, \beta) = \arg\min_{\gamma \in \Gamma_{\alpha,\beta}} \sum_{i=1}^{n} \sum_{j=1}^{M} \gamma_{i,j} \, \|i - j\|_2^2, \tag{5}$$

TABLE I: Datasets for training, validation, and evaluation.

Use	Dataset	Split	Dur.	Utt.	Spk.	Phone Time.
Train/Val.	LibriSpeech [26]	Train+Dev	971.6h	286.8k	2411	MFA-based[3]
Eval.	TIMIT [27]	Dev/Test	0.50h	400/192	50/24	Expert
	Buckeye [28]	Tr/Dev/Test	18.51h	7946	40	Expert
	L2-ARCTIC [29]	Test	3.66h	900	6	Expert
	SO762 [30]	Test	5.58h	2500	125	–

where $\Gamma_{\alpha,\beta} = \left\{\gamma \in \mathbb{R}_+^{n \times M} \mid \gamma \mathbf{1}_M = \alpha, \, \gamma^\top \mathbf{1}_n = \beta\right\}$ denotes valid couplings preserving the marginal distributions. Based on the *sequence optimal transport distance* framework [16], the OTTC loss becomes:

$$\mathcal{L}_{\mathrm{OTTC}} = -\sum_{i=1}^{n} \sum_{j=1}^{M} \gamma_{i,j}(\alpha, \beta) \cdot \log p(z_j \mid x_i), \tag{6}$$

where $p(z_j \mid x_i)$ is the posterior probability of the j-th topology state. As shown in Fig. 3, this transport process assigns probability mass to all topological states and produces a more complete subphonetic trace.

D. Topology-aware Decoding

At inference time, the topology is applied as a finite-state decoding constraint:

$$\begin{aligned} \hat{\mathbf{s}} &= \arg\max_{\mathbf{s} \in \mathcal{A}_\tau} \sum_{i=1}^{n} \log p(s_i \mid x_i), \\ \hat{\mathbf{Z}} &= \mathrm{collapse}_\tau(\hat{\mathbf{s}}), \qquad \hat{\mathbf{Y}} = \kappa_\tau(\hat{\mathbf{Z}}). \end{aligned} \tag{7}$$

where $\mathcal{A}_\tau$ is the set of valid frame-level state paths under the chosen topology. The decoded path is first collapsed into a subphonetic-state sequence $\hat{\mathbf{Z}}$, and $\kappa_\tau(\cdot)$ then maps this valid topology-state sequence back to phone labels. This distinction is important: an isolated substate is not sufficient to emit a phone. The decoded path must follow the state-transition rules illustrated in Fig. 2.

In TDFA, the reference monophone sequence $\mathbf{Y}$ is expanded into a topology-constrained graph, and a finite-state aligner[2] recovers the best path from the posteriors [25]. In the TI setting, no reference graph is supplied; we decode under the topology constraints and collapse only valid paths back to phone labels and boundaries.

IV. Experimental Evaluation

We evaluate the proposed acoustic model from three perspectives: segmentation accuracy, acoustic fidelity, and downstream PA performance.

A. Datasets and Experimental Setup

1) Datasets: As summarized in Table I, LibriSpeech [26] is the only corpus used for training the proposed acoustic models. Specifically, TIMIT [27], Buckeye [28], and L2-ARCTIC [29] provide human-annotated phone-level timestamps for segmentation and forced alignment evaluation, while speechocean762 (SO762) [30] is used for pronunciation-related evaluation. All phone labels are mapped to a 39-phone CMUdict inventory.

[2]Implemented with K2-FSA: https://github.com/k2-fsa/k2.

[3]Source: https://huggingface.co/datasets/gilkeyio/librispeech-alignments.

TABLE II: Baselines and proposed acoustic model's segmentation and recognition performance. Cano. and Perc. denote the canonical and perceived phone labels used in L2 corpora, respectively.

Model	Segmentation							Recognition / MDD								
	Text-dependent FA				Text-independent			PER↓							MDD F1↑	
	TIMIT Dev		TIMIT Test		R-value↑			TIMIT		Buckeye	L2-ARCTIC		SO762		L2-ARCTIC	SO762
	TSE↓	R-value↑	TSE↓	R-value↑	TIMIT Test	Buckeye	L2-ARCTIC	Dev.	Test		Cano.	Perc.	Cano.	Perc.		
Kaldi TDNN-F [31]	45.39	63.59	45.41	63.53	–	–	–	–	–	–	–	–	–	–	–	–
ZIPA† [32]	79.21	55.16	78.93	56.51	–	53.71	53.72	23.32	26.42	32.62	**12.03**	**19.79**	44.47	44.11	0.377	0.147
Charsiu FC–20ms [18]	40.72	76.55	40.36	76.38	**76.28**	70.02	75.99	22.04	24.17	31.83	14.68	21.73	43.93	43.49	0.431	0.188
Charsiu FC–10ms [18]	42.47	75.03	42.59	74.46	74.73	68.53	74.30	23.33	26.36	35.13	15.77	23.18	47.40	47.19	0.430	0.204
Ours–20ms	38.01	76.35	**37.40**	76.46	74.87	73.82	78.28	**20.79**	**23.20**	**29.74**	12.40	20.53	**39.77**	**39.41**	0.424	**0.240**
Ours–10ms	**36.37**	**77.93**	38.12	**78.16**	75.38	**74.09**	**78.94**	23.89	26.89	36.53	20.81	27.96	64.76	64.74	**0.486**	0.236
MFA 3.0 [33]	30.73	87.83	31.95	87.44	–	–	–	–	–	–	–	–	–	–	–	–

†ZIPA originally uses a 127-symbol IPA inventory. For a fair comparison, we decoded IPA sequences and mapped them to the 39-symbol ARPAbet inventory used in this study.

TABLE III: Ablation on L2-ARCTIC. All models were trained from scratch on the L2-ARCTIC training set, following common practice for MDD. Notably, the CTC conditioning head in Fig. 1 was not applied in this ablation.

(A) Effect of training criterion at 20 ms.

Objective	Topology	TD FA		TI seg.	Recog. / MDD	
		F1↑	TSE↓	R-val↑	PER↓	MDD F1↑
CTC	S1T1	32.73	82.90	42.58	18.19	0.562
CTC	S3T2⋆⋆	38.61	78.14	47.60	**16.68**	0.589
LPCTC [17]	S1T1	49.22	69.32	56.65	18.07	0.610
OTTC	S1T1	57.02	46.65	63.31	18.44	0.604
OTTC	S3T2⋆⋆	**62.91**	**40.64**	**68.34**	19.24	**0.619**

(B) Topology effect at different frame resolutions.

Topology	20 ms				10 ms			
	TSE↓	R-value↑	PER↓	MDD↑	TSE↓	R-value↑	PER↓	MDD↑
S1T1	46.65	63.31	**18.44**	0.604	48.09	66.46	**19.87**	0.632
S2T1	39.99	68.38	18.79	0.618	41.49	69.72	20.89	0.641
S2T1⋆	**39.53**	**68.72**	18.91	**0.623**	42.42	69.55	20.59	0.644
S3T2	40.94	67.77	19.27	0.617	39.78	71.17	22.18	**0.651**
S3T2⋆	40.72	68.40	19.00	0.616	39.59	71.21	21.43	0.644
S3T2⋆⋆	40.64	68.34	19.24	0.619	**39.25**	**71.49**	21.14	0.645

2) Model Design: As shown in Fig. 1, the proposed model consists of a trainable SSL encoder followed by two prediction heads. The CTC head learns the monophone sequence $\mathbf{Y}$, while the SP head learns the expanded subphonetic-state sequence $\mathbf{Z}$ with OTTC. Specifically, we use the CTC posterior to condition the SP head through FiLM-style linear modulation [34]. For each frame i, let $\mathbf{p}_i^{\text{ctc}}$ denote the CTC posterior and $\mathbf{h}_i^{\text{sp}}$ denote the hidden representation of the SP head. The conditioned SP representation is computed as

$$\widetilde{\mathbf{h}}_i^{\text{sp}} = \big(1 + g(\mathbf{p}_i^{\text{ctc}})\big) \odot \mathbf{h}_i^{\text{sp}} + b(\mathbf{p}_i^{\text{ctc}}), \tag{8}$$

where $g(\cdot)$ and $b(\cdot)$ are linear projections and $\odot$ denotes element-wise multiplication. The overall training objective combines the CTC loss and the topology-aware OTTC loss:

$$\mathcal{L} = \lambda \mathcal{L}_{\text{CTC}} + (1-\lambda)\mathcal{L}_{\text{OTTC}}, \tag{9}$$

where $\lambda = 0.5$ in our setting.

3) Training Setup: We trained two frame-rate variants of the proposed model. Ours–20ms uses the standard WavLM-large [35] with output resolution at 20 ms, while Ours–10ms changes the final convolutional stride to obtain 10 ms frame-level outputs. The encoder and both prediction heads are trained jointly for up to 50 epochs with a batch size of 32 using only phone labels and the objective in Eq. (9), followed by 1000 steps of frame-wise CE supervision to stabilize silence frames, as in Charsiu [18]. Each prediction head is a two-layer MLP with hidden size 384 and ReLU activation. Checkpoints are selected by LibriSpeech development PER. Based on Sec. IV-D, Ours–20ms uses S2T1, while Ours–10ms uses $\text{S3T2}\star\star$ in the main experiments.

B. Evaluation Metrics

1) Segmentation Accuracy: For TDFA, we report time-stamp error (TSE) [36]:

$$\text{TSE} = \frac{1}{N}\sum_{i=1}^{N}\left(|\hat{s}_i - s_i| + |\hat{e}_i - e_i|\right), \tag{10}$$

where (s_i, e_i) and $(\hat{s}_i, \hat{e}_i)$ denote the reference and predicted start and end times of the i-th phone segment, respectively, and N is the number of evaluated phone segments. For TI segmentation, the predicted phone sequence may have a different length. For these methods, we report the R-value by treating segmentation as boundary detection over timestamp sets following [37], [38]:

$$\begin{aligned} P_{\text{seg}} &= \frac{|\mathcal{M}_\delta|}{|\widehat{\mathcal{T}}|}, \qquad R_{\text{seg}} = \frac{|\mathcal{M}_\delta|}{|\mathcal{T}|}, \qquad \text{OS} = \frac{R_{\text{seg}}}{P_{\text{seg}}} - 1, \\ r_1 &= \sqrt{(1-R_{\text{seg}})^2 + \text{OS}^2}, \qquad r_2 = \frac{-\text{OS} + R_{\text{seg}} - 1}{\sqrt{2}}, \\ &\text{R-value} = 1 - \frac{|r_1| + |r_2|}{2}, \end{aligned} \tag{11}$$

where boundary-matching tolerance δ is set to 20 ms.

2) Recognition and Diagnosis Accuracy: For phone recognition, we report phone error rate (PER):

$$\text{PER} = \frac{S + D + I}{N}, \tag{12}$$

where S, D, and I are phone substitutions, deletions, and insertions, and N is the number of reference phones. On L2-ARCTIC and SO762, we compute two types of PER: canonical PER measures dictionary-derived recognition, while perceived PER tests fidelity to the learner's realized pronunciation beyond canonical bias. We also report MDD F1 following the standard detection protocol [39]:

$$\begin{aligned} P_{\text{MDD}} &= \frac{\text{TR}}{\text{TR} + \text{FR}}, \quad R_{\text{MDD}} = \frac{\text{TR}}{\text{TR} + \text{FA}}, \\ F1_{\text{MDD}} &= \frac{2P_{\text{MDD}}R_{\text{MDD}}}{P_{\text{MDD}} + R_{\text{MDD}}}, \end{aligned} \tag{13}$$

where TR, FR, and FA denote true rejection (detected mispronunciation), false rejection, and false acceptance in MDD, respectively.

3) Pronunciation Assessment Performance: For APA, we evaluate phone-level mean squared error (MSE) and Pearson correlation coefficient (PCC) on SO762 following [5]. MSE measures the prediction error of phone-level pronunciation scores, while PCC reflects the agreement between predicted scores and human annotations. Completeness at the utterance level is omitted for compactness.

C. Segmentation and Recognition Performance

Table II summarizes the main results for text-dependent forced alignment (TDFA), text-independent (TI) segmentation/recognition, and mispronunciation detection.

For TDFA, MFA remains the strongest topline, consistent with prior forced-alignment comparisons [40]. Among neural non-MFA systems, the proposed models give the best overall segmentation trade-off: Ours–10ms obtains the highest TD R-values on both TIMIT splits and the best TI R-values on Buckeye and L2-ARCTIC, while Ours–20ms gives the lowest non-MFA TSE on TIMIT test.

The recognition and MDD results show that the segmentation gain does not come from sacrificing phone-level evidence. Ours–20ms obtains the lowest PER on TIMIT dev/test, Buckeye, and both canonical/perceived SO762 labels, and remains second-best on both L2-ARCTIC label sets after ZIPA. Meanwhile, the proposed models achieve the best MDD F1 on both L2-ARCTIC and SO762. Overall, these results suggest that the proposed model improves temporal localization while preserving robust acoustic evidence for recognition and downstream pronunciation-related tasks.

D. Ablation Studies

To isolate the contribution of each component in Sec. III, we trained the ablation models from scratch on L2-ARCTIC, following common MDD practice.

1) Effect of Training Criterion: Table III-A separates the effects of topology and training criterion. Under CTC, the richer S3T2⋆⋆ topology gives only modest segmentation gains, suggesting that topology alone does not ensure meaningful phone-internal state traversal. The main improvement comes from replacing CTC with OTTC under the same topology, which improves TD F1 from 38.61 to 62.91, reduces TSE from 78.14 ms to 40.64 ms, and raises TI R-value from 47.60 to 68.34. LPCTC [17] provides an additional CTC variation as mentioned in Sec. II-A, but remains below OTTC in segmentation. Although CTC+S3T2⋆⋆ has the lowest PER, OTTC achieves the best MDD F1, indicating better temporal localization without losing pronunciation-relevant acoustic evidence.

2) Effect of Subphonetic Topologies: Table III-B shows the effect of subphonetic topology at different frame resolutions.

Segmentation. Finer frame resolution better supports subphonetic state transitions. At 10 ms, increasing the topology resolution yields a clear boundary gain, improving the R-value by 5.03 points and reducing TSE by 8.84 ms compared with the monophone topology. At 20 ms, the maximum gain appears at S2, while S3 topologies do not provide significant additional benefits. This suggests that 20 ms frames are often too coarse to resolve three-state phone-internal transitions, whereas 10 ms frames provide more temporal detail for longer subphonetic modeling.

Recognition and MDD. Table III-B also shows a trade-off between recognition and diagnosis. PER is generally better at 20 ms, where the model makes fewer frame-level decisions for sequence recognition. MDD follows the opposite trend: the best 10 ms topology improves MDD F1 by an absolute 0.03 over the best 20 ms setting.

TABLE IV: Comparison with representative APA systems GOPT [5] and ConPCO [41]. Indented rows denote controlled variants that use the same scoring pipeline but replace either the frame-level log posteriors or the phone-level timestamps. Acc., Str., Flu., Pro., and Tot. denote accuracy, stress, fluency, prosody, and total score, respectively.

Model	Phone		Word PCC			Utt. PCC			
	MSE ↓	PCC ↑	Acc. ↑	Str. ↑	Tot. ↑	Acc. ↑	Flu. ↑	Pro. ↑	Tot. ↑
Kaldi TDNN-F [5]	0.085	0.612	0.533	0.291	0.549	**0.714**	**0.753**	**0.760**	**0.742**
w/ Charsiu's LogP	0.126	0.293	0.308	0.094	0.317	0.585	0.661	0.660	0.614
w/ Ours LogP	0.130	0.272	0.297	0.040	0.306	0.580	0.626	0.635	0.596
w/ Charsiu's TS	0.124	0.300	0.306	0.189	0.312	0.517	0.596	0.583	0.536
w/ Ours TS	0.123	0.302	0.307	0.277	0.310	0.513	0.592	0.580	0.531
Charsiu TS&LogP	0.101	0.518	0.464	0.151	0.476	0.682	0.705	0.705	0.701
Ours TS&LogP	**0.084**	**0.617**	**0.565**	**0.382**	**0.578**	0.712	0.712	0.717	0.723
HierCB+ConPCO [41]	**0.071**	**0.701**	**0.669**	**0.437**	**0.682**	0.780	0.830	0.823	0.803
w/ Charsiu's TS	0.083	0.635	0.589	0.221	0.602	0.785	0.835	0.831	0.812
w/ Ours TS	0.080	0.658	0.621	0.215	0.633	**0.796**	**0.837**	**0.833**	**0.819**

E. Downstream Task: Pronunciation Assessment Performance

We further examine whether improved acoustic evidence and alignment benefit downstream APA. Table IV reports the results on SO762 [30] using representative APA systems. For the GOP-NN-based APA pipeline, i.e., GOPT [5] in our setting, the input features depend on both frame-wise acoustic confidence and phone-level segmentation accuracy [42]. To disentangle posterior quality from alignment quality, we evaluate three GOPT-based settings: (i) Fixed FA, where Kaldi-TDNN-F timestamps are fixed and LogP is varied across acoustic models; (ii) Fixed LogP, where Kaldi-TDNN-F LogP is fixed and timestamps are varied across aligners; and (iii) Full model, where each acoustic model provides both LogP and timestamps.

Replacing only one component generally hurts the GOPT pipeline, suggesting that mismatched LogP and timestamp sources disrupt GOP feature calibration. In the full setting, however, our model outperforms Charsiu and the original GOPT at the phone and word levels, achieving the best phone-level MSE/PCC and improving word-level Stress PCC to 0.382. This confirms that finer temporal granularity and refined posterior modeling improve localized pronunciation feedback. At the utterance level, the GOPT-based pipeline still lags behind the Kaldi TDNN-F baseline. We attribute this to the limited long-range information provided by GOP-style LogP features, which mainly capture local phoneme-level acoustic confidence but are less suited to modeling rhythm, speaking rate, pauses, and prosodic fluency. This is consistent

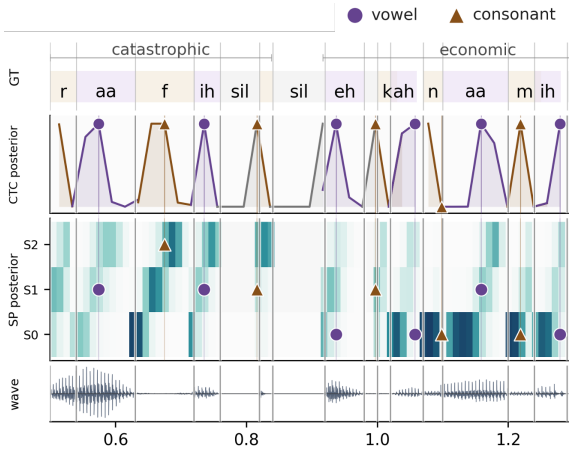
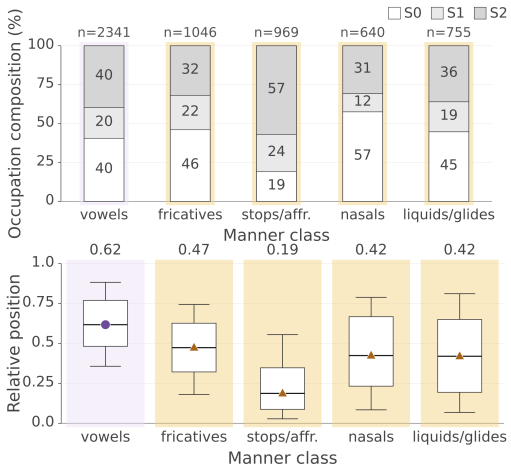


Fig. 4: CTC peaks as acoustic landmarks for probing subphonetic states



(a) Articulation-Manner-level S3 occupancy and CTC peak centers.

Group	Vowel.	Fricatives.	Stop/aff.	Nasals.	Liquids.
phones	*aa ae ih uw*	*f s sh z*	*b d k ch*	*m n ng*	*l r w y*

(b) Phones grouped by the manner of articulation.

Fig. 5: Subphonetic topology and CTC-probe analysis for pronunciation modeling.

with prior work reporting a trade-off between local and global pronunciation scoring [43].

The HierCB+ConPCO [41] experiments examine the effect of segmentation accuracy when contextual features are also modeled. Unlike the GOPT-based setting, which uses only GOP-NN features, HierCB+ConPCO combines GOP-NN features with segmented SSL representations. Here, GOP-NN features provide local pronunciation evidence, while SSL features add contextual and prosodic cues for holistic scoring. As shown in Table IV, the utterance-level gains obtained with our timestamps suggest that more accurate boundaries help the hierarchical scorer better exploit SSL representations for modeling fluency, prosody, and overall pronunciation quality.

V. Subphonetic Acoustic Modeling and Phoneme-Dependent Characteristics

To demonstrate that the proposed SP states encode acoustic structure rather than arbitrary frame-level partitions, we analyze the Ours–10ms model from two complementary perspectives. First, we compute the S3 state occupancy for each phone and then aggregate phones by manner group, with representative phones summarized in Fig. 5(b). Fig. 5(a) shows that vowels, fricatives, and liquids/glides have relatively balanced state distributions, whereas stops/affricates and nasals behave differently: stops/affricates place a 57% share on S2, while nasals are dominated by S0. This suggests that the learned topology is not a uniform three-way partition across phonemes, but reflects manner-dependent acoustic structure.

Second, we use the CTC conditioning head shown in Fig. 1 as a probe. Specifically, across five random seeds, we train the CTC conditioning branch to a comparable recognition level (LibriSpeech dev PER below 4.5%) and locate the frame with the maximum non-blank CTC posterior within each reference phone interval. The relative position is normalized to $[0, 1]$, where 0 and 1 denote the first and last frames of the phone interval. As illustrated in Fig. 4, this peak provides an acoustic landmark for asking where the most discriminative evidence falls inside the ordered S3 topology.

The CTC probe reveals a clear vowel–consonant contrast. For vowels, the average CTC peak position occurs later in the phone interval, around 0.62. In contrast, consonants show earlier peak positions. Notably, stops/affricates have the earliest peak center, around 0.19, which is consistent with their strong onset or release cues; phones such as /b/ and /k/ concentrate much of their discriminative evidence near the beginning of the phone interval. These observations align with previous research [44], [45] and acoustic phonetics [46]: vowels are typically characterized by relatively stable central steady-state regions, whereas consonants often rely on short transient cues such as closure, release, or onset transitions.

The learned topology therefore does more than improve metrics: it captures phoneme-dependent acoustic structure below the phone level. This also suggests a possible diagnostic use beyond the present experiments. For example, vowel quality could be measured around nucleus-like central or later states, while consonantal errors could be examined around onset-, release-, or transition-related states, yielding a content-aware GOP-like feature for PA.

VI. Conclusion and Future Work

This paper presents a subphonetic acoustic modeling framework that bridges HMM-inspired ordered state modeling and end-to-end alignment learning through OTTC. Experiments across read, spontaneous, and L2 speech demonstrate improvements in forced alignment and text-independent segmentation while maintaining competitive recognition performance, with further gains in downstream APA. Ablation studies and probing show that the learned states capture meaningful phone-internal acoustic structure. In future work, we will further investigate how these subphonetic posteriors can be used for more interpretable pronunciation measurements.

REFERENCES

[1] L. Davis and J. M. Norris, *Challenges and Innovations in Speaking Assessment*. Routledge, 2025.

[2] S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2, pp. 95–108, 2000.

[3] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in L2 English speech using multidistribution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 25, no. 1, pp. 193–207, 2017.

[4] W.-K. Leung, X. Liu, and H. Meng, "CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 8132–8136.

[5] Y. Gong, Z. Chen, I.-H. Chu, P. Chang, and J. Glass, "Transformer-based multi-aspect multi-granularity non-native English speaker pronunciation assessment," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7262–7266.

[6] X. Cao, Z. Fan, T. Svendsen, and G. Salvi, "A framework for phoneme-level pronunciation assessment using CTC," in *Proc. of Interspeech*, 2024, pp. 302–306.

[7] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[8] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The HTK book," *Cambridge university engineering department*, vol. 3, no. 175, p. 12, 2002.

[9] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Proc. of Interspeech*, 2017, pp. 498–502.

[10] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. of the international conference on Machine learning (ICML)*, 2006, pp. 369–376.

[11] H. Geng, L. Yang, X. Chen, H. Sun, D. Saito, and N. Minematsu, "Beyond acoustic sparsity and linguistic bias: A prompt-free paradigm for mispronunciation detection and diagnosis," *arXiv preprint arXiv:2604.22133*, 2026.

[12] B. Lin and L. Wang, "Phoneme mispronunciation detection by jointly learning to align," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6822–6826.

[13] A. Zeyer, R. Schlüter, and H. Ney, "Why does CTC result in peaky behavior?" *arXiv preprint arXiv:2105.14849*, 2021.

[14] J. Tian, B. Yan, J. Yu, C. Weng, D. Yu, and S. Watanabe, "Bayes risk CTC: Controllable CTC alignment in sequence-to-sequence tasks," *arXiv preprint arXiv:2210.07499*, 2022.

[15] Z. Yao, W. Kang, F. Kuang, L. Guo, X. Yang, Y. Yang, L. Lin, and D. Povey, "Delay-penalized CTC implemented based on finite state transducer," in *Proc. Interspeech*, 2023, pp. 1329–1333.

[16] Y. Kaloga, S. Kumar, P. Motlicek, and I. Kodrasi, "A differentiable alignment framework for sequence-to-sequence modeling via optimal transport," *arXiv preprint arXiv:2502.01588*, 2025.

[17] R. Huang, X. Zhang, Z. Ni, L. Sun, M. Hira, J. Hwang, V. Manohar, V. Pratap, M. Wiesner, S. Watanabe, D. Povey, and S. Khudanpur, "Less peaky and more accurate CTC forced alignment by label priors," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 831–11 835.

[18] J. Zhu, C. Zhang, and D. Jurgens, "Phone-to-audio alignment without text: A semi-supervised approach," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8167–8171.

[19] K. J. Shih, R. Valle, R. Badlani, A. Lancucki, W. Ping, and B. Catanzaro, "RAD-TTS: Parallel flow-based TTS with robust alignment learning and diverse synthesis," in *Proc. of the International Conference on Machine Learning Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021.

[20] K.-F. Lee, H.-W. Hon, M.-Y. Hwang, S. Mahajan, and R. Reddy, "The SPHINX speech recognition system," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1989, pp. 445–448 vol.1.

[21] L. Bahl, J. Bellegarda, P. deSouza, P. Gopalakrishnan, D. Nahamoo, and M. Picheny, "A new class of fenonic Markov word models for large vocabulary continuous speech recognition," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1991, pp. 177–180 vol.1.

[22] M. Hwang and X. Huang, "Subphonetic modeling with markov states-senone," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 1992, pp. 33–36 vol.1.

[23] Z. Zhao and P. Bell, "Advancing CTC models for better speech alignment: A topological approach," in *Proc. of IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 279–285.

[24] Z. Zhao, P. Chen, and P. Bell, "Regarding topology and adaptability in differentiable WFST-based E2E ASR," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024, pp. 843–847.

[25] D. Povey, P. Zelasko, and S. Khudanpur, "Speech recognition with next-generation Kaldi (k2, lhotse, icefall)," in *Proc. of Interspeech: Tutorials*, 2021.

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

[27] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

[28] M. A. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymond, "The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability," *Speech Communication*, vol. 45, no. 1, pp. 89–95, 2005.

[29] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-ARCTIC: A non-native English speech corpus," in *Proc. of Interspeech*, 2018, p. 2783–2787.

[30] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang, "speechocean762: An open-source non-native English speech corpus for pronunciation assessment," in *Proc. of Interspeech*, 2021.

[31] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. of Interspeech*, 2018, pp. 3743–3747.

[32] J. Zhu, F. Samir, E. Chodroff, and D. R. Mortensen, "ZIPA: A family of efficient models for multilingual phone recognition," in *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds., 2025.

[33] M. McAuliffe, K. Gunter, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner and the state of speech-to-text alignment in 2026," *arXiv preprint arXiv:2606.18466*, 2026.

[34] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: visual reasoning with a general conditioning layer," in *Proc. of the Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.

[35] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing (JSTSP)*, vol. 16, no. 6, pp. 1505–1518, 2022.

[36] X. Zhang, V. Manohar, D. Zhang, F. Zhang, Y. Shi, N. Singhal, J. Chan, F. Peng, Y. Saraf, and M. Seltzer, "On lattice-free boosted mmi training of hmm and ctc-based full-context asr models," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 1026–1033.

[37] L. Strgar and D. Harwath, "Phoneme segmentation using self-supervised speech models," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 1067–1073.

[38] F. Kreuk, J. Keshet, and Y. Adi, "Self-supervised contrastive learning for unsupervised phoneme segmentation," in *Proc. of Interspeech*, 2020, pp. 3700–3704.

[39] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in L2 English speech using multidistribution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 25, no. 1, pp. 193–207, 2016.

[40] R. Rousso, E. Cohen, J. Keshet, and E. Chodroff, "Tradition or innovation: A comparison of modern ASR methods for forced alignment," in *Proc. Interspeech*, 2024, pp. 1525–1529.

[41] B.-C. Yan, Y.-C. Wang, J.-T. Li, M.-S. Lin, H.-W. Wang, W.-C. Chao, and B. Chen, "ConPCO: Preserving phoneme characteristics for automatic pronunciation assessment leveraging contrastive ordinal regularization," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[42] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154–166, 2015.

[43] J.-T. Li, B.-C. Yan, Y.-C. Wang, and B. Chen, "Multi-task pretraining for enhancing interpretable l2 pronunciation assessment," in *Proc. of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2025, pp. 531–536.

[44] T. Wu, J. Duchateau, and D. Van Compernolle, "Phoneme dependent frame selection preference." in *Proc. of Interspeech*, 2007, pp. 50–53.

[45] T. Wu, D. Van Compernolle, J. Duchateau, and H. Van Hamme, "Maximum likelihood based temporal frame selection," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 2006, pp. I–I.

[46] R. Wayland, *Phonetics: A practical introduction*. Cambridge University Press, 2018.