Multi-Task Learning for Non-Canonical Phoneme Recognition via Articulatory Feature Decomposition

Sophia Riaz*[1],Haoze Zheng*[,1], Amos Roche[1], Miyu Zhang[1], Anamika Ragu[1], Salvatore Penachio[1], Kaustav Mukherjee[1], Aneesh Jonelagadda**[1]

Kaliber AI, San Mateo CA

Pathological and more broadly non-canonical speech present significant challenges for automatic phoneme recognition due to systematic deviations from canonical pronunciation and limited availability of labeled clinical speech data. Existing phoneme recognition systems are typically trained on canonical speech and treat phonemes as atomic categorical labels, limiting their ability to detect structured articulatory errors common in speech disorders and accents. In this work, we introduce a linguistically structured approach to non-canonical phoneme recognition that decomposes phoneme prediction into articulatory feature dimensions such as manner, place, and voicing. We implement this formulation using a hierarchical multi-task learning architecture in which task-specific articulatory feature heads learn feature-level representations that are subsequently integrated through a cross-attention-based fusion module to produce phoneme predictions. To address the scarcity and noise of pathological speech labels, we combine this framework with semi-supervised learning via Momentum Pseudo-Labeling (MPL) and propose a cascaded training strategy that progressively introduces articulatory feature tasks while employing staged unfreezing of a pretrained speech encoder. Experiments on L2-ARCTIC, used as a proxy for pathological speech variation, show that the proposed approach achieves substantial improvements in phoneme recognition performance compared to strong baseline architectures, while yielding interpretable error patterns aligned with phonological feature structure. These results suggest that articulatory feature supervision is a promising strategy for robust and interpretable phoneme recognition in non-canonical speech, and motivate future validation on clinically diagnosed pathological speech datasets.

1. Introduction

Spoken language is the primary modality of human communication. Through speech, humans convey not only linguistic content, but also emotion, context, and even indicators of health. When speech production is impaired, often to the extent of a disorder, this communication becomes degraded, reducing the effectiveness of such conveyance. As public awareness of speech disorders has increased, so too has the demand for clinical intervention (American Speech-Language-Hearing Association 2023), including

* Equal contribution
** Corresponding author

at-home speech therapy which entails immediate and detailed feedback on speech production. This requires systems that not only detect incorrect utterances but also provide fine-grained error analysis. Such analysis can be performed at both the phoneme and prosodic levels. While prosodic analysis is relatively robust in modern audio processing pipelines (Cohen et al. 2025), phoneme-level analysis remains a significant challenge.

Pathological phoneme recognition has substantial room for improvement. Existing phoneme recognition models are primarily trained on healthy speech, and thus struggle to generalize to disordered speech (Berisha and Liss 2024). Models that incorporate pathological data typically have limited coverage of the spectrum of disorders. As a result, performance on common conditions such as Motor Speech Disorder (MSD) and Childhood Apraxia of Speech (CAS) remains poor, with reported Phoneme Error Rates (PER) ranging from 42% to 69% (Li et al. 2020; Ravanelli et al. 2021; Baevski et al. 2020). When trained exclusively on healthy speech, these models exhibit a tendency to normalize atypical pronunciations to their closest canonical forms. For example, an utterance of "wapit" (/ẅæpɪt/) may be incorrectly mapped to the canonical phoneme sequence for "rabbit" (/ɹ̈æbɪt/). This auto-corrective bias arises because pathological pronunciations are not represented in the training data, highlighting the need for models explicitly trained on non-canonical speech. However, training on pathological speech is fundamentally constrained by the scarcity of high-quality labeled data (Rudzicz, Namasivayam, and Wolff 2012; Kim et al. 2008; Menéndez-Pidal et al. 1996) with ground-truth annotations that may themselves be noisy due to inter-annotator variability (Strömbergsson et al. 2020). Momentum pseudo-labeling (MPL) (Yang et al. 2022) provides a natural approach to mitigating this through a teacher–student framework that leverages unlabeled data.

Despite advances in self-supervised speech models such as Wav2Vec2 and WavLM (Baevski et al. 2020; Chen et al. 2022), existing models treat phonemes as atomic categorical labels and do not explicitly encode phonological structure. While effective for canonical speech, this is poorly suited to non-canonical speech where errors are often systematic: speakers may preserve certain articulatory properties while deviating along others (Duffy 2019). Articulatory features provide a linguistically motivated decomposition of phonemes that correspond to physical properties of speech production.

To address these challenges, we propose a hierarchical articulatory multi-task learning framework for low-resource non-canonical phoneme recognition that jointly predicts articulatory features and phonemes using cross-attention-based feature fusion. We further incorporate MPL and speech-specific data augmentation to improve robustness. We evaluate our approach on L2-ARCTIC using Phoneme Error Rate (PER), mispronunciation detection metrics, and ablation studies. Beyond overall performance, we analyze error patterns along articulatory feature dimensions. Our results demonstrate consistent improvements in PER over strong baselines and highlight the effectiveness or structured articulatory supervision for low-resource non-canonical settings.

We therefore investigate the following research questions:

1. To what extent can articulatory structure be incorporated into phoneme recognition through hierarchical articulatory feature supervision?
2. Do phoneme recognition errors exhibit structure along articulatory feature dimensions, and can modeling these dimensions reduce such errors?
3. Can our novel architecture trained on limited non-canonical speech detect non-canonical speech patterns better than baseline models trained on a much larger corpus of canonical speech?

1.1 Main Contributions

Our main contributions are as follows:

- We introduce a hierarchical multi-task learning (HMTL) architecture, including heuristically-motivated augmentations, for non-canonical phoneme recognition that decomposes phoneme prediction into articulatory feature dimensions integrated through a cross-attention–based fusion layer to produce final phoneme predictions.
- We propose a structured articulatory supervision framework that derives rich auxiliary labels from ARPAbet and exploits linguistic relationships between phonemes.
- We conduct a comprehensive empirical evaluation with detailed ablations, comparisons and analysis where we demonstrate that phoneme recognition errors exhibit systematic structure along articulatory features, motivating the HMTL architecture.

2. Background

2.1 Clinical and Linguistic Foundations of Motor Speech Disorders

Motor speech disorders are characterized by impairments in the processes that transform linguistic intent into coordinated speech output. These processes span multiple stages, from motor planning to the execution of articulatory movements. Two main types prevail in clinical literature: apraxia of speech and dysarthria.

Childhood apraxia of speech (CAS) is a rare neurological speech disorder in which the brain struggles to plan and sequence the complex muscle movements needed for speech despite presenting without neuromuscular deficits. This contributes to a wide range of symptoms, including distorted sounds, inconsistent errors in speech, groping for sounds, and incorrect prosody. These symptoms manifest in a wide array of phonetic errors. Unlike CAS, where motor planning is disrupted, dysarthric speech is the result of impaired motor execution. Six presentations arise from distinct underlying pathologies (Cleveland Clinic 2025) and can be broadly characterized in terms of linguistic impairments and alterations in speech quality. The breadth of these deviations from typical speech presents a significant challenge for present automatic speech recognition systems, thereby motivating the need for strategies that better capture the manifestations of this variability.

2.2 Phoneme Recognition for Non-Canonical Speech

Fundamentally, the development of Automatic Speech Recognition (ASR) models is geared towards achieving the lowest corpus-level word error rate (WER), underscoring that frontier systems are optimized for word-level transcription accuracy rather than for the preservation of fine-grained phonetic detail (Fatehi, Torres Torres, and Kucukyilmaz 2025). The same inductive biases that enable low WER, such as strong language modeling and distributional smoothing, also encourage normalization of disordered speech, mapping atypical pronunciations to the nearest valid phoneme. While this abstraction is beneficial for robust transcription of speech, it obscures the subtle deviations in articulation that are core to diagnosing and treating pathologies arising from motor speech

disorders. Transitioning from word-level ASR to phoneme recognition introduces a distinct set of modeling challenges that have important architectural implications. Contemporary ASR pipelines measure mispronunciation through detection of deviations relative to discrete word and subword level targets but they rarely characterize *how* mispronunciations manifest acoustically. At a phoneme level this implementation is not robust enough as clinical descriptions of disordered speech are inherently feature-based, emphasizing dimensions such as articulatory precision, prosody, and voicing control. Additionally, temporal resolution becomes significantly more critical and models must be able to capture fine-grained acoustic transitions corresponding to rapid articulatory movements. Unlike words or subwords, phonemes exhibit high context dependence (e.g. co-articulation effects), shorter durations, and greater acoustic overlap, necessitating architectures that can preserve local temporal detail while still modeling longer-range dependencies.

2.3 Pathology Data Scarcity and L2 as Proxy

The challenges of pathological ASR are compounded by the scarcity of high-quality labeled clinical speech data. Collecting and annotating such data is complicated, resulting in datasets that are orders of magnitude smaller than those used in standard ASR, and where inter-annotator variability can be substantial (Hosom 2009). This motivates the search for a proxy dataset that can minimally identify mispronunciations in its ground truth. Therefore, we adopt L2-ARCTIC, an accented corpus with aligned phonemic annotations. Although derived from non-native speech, it provides systematic pronunciation deviations while maintaining annotation quality, offering the most practical reconciliation between scale and supervision (Zhao et al. 2018). These typologically distinct non-native speakers introduce systematic phoneme variation, where L1 phonological constraints shape L2 articulation in a manner that is structurally comparable to certain pathological deviation patterns (Farish, Davis, and Wilson 2020; Flege 1995; Best and Tyler 2007). Namely, reduced vowel space, common among East Asian speakers, resembles the centralized vowel production observed in hypokinetic dysarthria (Kent and Kim 2003). These accent-driven deviations are predictable within each L1 group, but collectively produce a range of phonetic variability mirroring disordered speech. Nevertheless, while accented speech constitutes the most phonetically comprehensive proxy available, it does not represent the entire spectrum of variability observed in clinical populations, especially impairments caused by respiratory, phonatory, and prosodic systems (Duffy 2019). Thus, closing the gap requires strategies that can simultaneously leverage unlabeled data while diversifying the training dataset.

2.4 Articulatory Feature Representations

Phoneme production can be decomposed along articulatory dimensions that describe the configuration of the speech apparatus during sound generation. These categories vary between consonants and vowels. Key attributes for consonants include **place** of articulation, which specifies the location in the vocal tract where air flow is obstructed, **manner** of articulation which describes how the airflow is modified, and **voicing** which indicates if the vocal chords are vibrating. Vowels, in contrast, are characterized by tongue position and shape: **backness**, which represents tongue position, **height** which reflects the vertical tongue position, and **roundedness**, indicating lip roundness. Each phoneme can be represented as a combination of such features. Importantly, these features are not all mutually exclusive. Demonstrably, /r/ can occupy multiple place

categories according to dialect and individual (Ladefoged and Johnson 2014). Complementing these features, phonological structures operate more abstractly, encoding how phonemes sound within a language rather than their physiological catalysts. Features such as rhoticity encode patterns of linguistic behavior that may not map to an articulatory configuration. Together, phonological and articulatory features form a structured, multi-dimensional representation of speech. This structured representation suggests that phoneme recognition is more naturally formulated as a multi-label prediction problem rather than as classification over independent atomic labels. Furthermore, articulatory features exhibit hierarchical dependencies as vowel and consonant categories constrain the set of valid feature combinations.

2.5 Related Work

The most direct precursor to this work is authored by Yang et al. (2022), who propose a mispronunciation detection framework built on Wav2Vec2.0 and MPL for accented (L2) speech. Their approach demonstrates that combining self-supervised acoustic representations with a teacher—student semi-supervised training regime achieves improvements over purely supervised baselines, motivating our adoption of MPL for low-resource non-canonical speech. However, several limitations remain in the aforementioned work (Yang et al. 2022). The formulation remains largely phoneme-centric and as a sequence of atomic phoneme labels. Thus, deviations are reduced to correctness scores without capturing the underlying articulatory structure of errors. Furthermore, robustness to non-canonical acoustic variability relies primarily on the pretrained backbone, whereas our framework incorporates speech-specific data augmentation to improve generalization under diverse pronunciation patterns.

Existing literature explores phonological and articulatory features. Classical approaches model speech as binary compositions of articulatory features instead of discrete labels. Early approaches modeled speech as binary articulatory compositions (Stouten and Martens 2006), while Mortensen et al. (2016) provides linguistically motivated feature inventories. More recent work has revisited these representations with deep learning. Weakly supervised phonological bottlenecks have been used for intelligibility prediction (Thienpondt et al. 2025), while Phonet estimates frame-level phonological posteriors (Tadavarthy, Jones, and Renwick 2024). Multilingual speech modeling has benefited from phonological representations (Xu, Baevski, and Auli 2022), which have also been directly supervised for sequence prediction (Shahin, Epps, and Ahmed 2025). Despite these advances, articulatory information is primarily employed as an intermediate representation, bottleneck, or downstream feature to improve or analyze speech representations, rather than serving as the central inductive bias for phoneme recognition.

Multi-task learning (MTL) has been widely adopted to improve phoneme recognition and mispronunciation detection by jointly optimizing auxiliary objectives alongside a primary task. EL Kheir, Chowdhury, and Ali (2023) propose a multi-view multi-task framework that combines multilingual and monolingual encoders with auxiliary prediction of place and manner, improving phoneme recognition in low-resource settings. However, the auxiliary supervision is implemented as parallel classification heads and is limited to a subset of articulatory categories. Similarly, Pang et al. (2026) introduce a hierarchical MTL framework for pronunciation assessment, incorporating attention mechanisms to model correlations across linguistic levels. While hierarchical, the supervision remains focused on scoring rather than explicit phoneme-feature decomposition. Glocker and Georges (2024) propose a hierarchical MTL framework for

cross-lingual phoneme recognition by conditioning phoneme prediction on articulatory attribute predictions. Their work focuses on multilingual transfer using binary articulatory attributes and propagates probability distributions to the phoneme classifier. In contrast, we supervise a set of articulatory categories, integrating learned representations through cross-attention prior to phoneme decoding.

Collectively, these demonstrate that structured auxiliary supervision to improve learning is promising. However, current approaches either employ parallel supervision, target pronunciation or multilingual transfer, or condition prediction on non-articulatory features or probabilities. There remains limited work on hierarchical articulatory supervision for non-canonical phoneme recognition, grounded in the underlying physiology of speech production in low-resource conditions.

3. Methods

3.1 Articulatory Feature Derivation

WavLM is pre-trained on the Librispeech corpus (Panayotov et al. 2015), and further fine-tuned using L2-ARCTIC, which both primarily adopt ARPAbet phonemic transcriptions (Zhao et al. 2018). While ARPAbet provides a compact representation of speech, standard articulatory feature mappings are typically defined over IPA (International Phonetic Association 1999; Carnegie Mellon University 1998; Ladefoged and Johnson 2014). Consequently, the phoneme labels available for training do not explicitly

IPA Consonants (pulmonic)		Place										
		Labial		Coronal					Dorsal			Laryngeal	
		Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Manner	Plosive / Stop	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
	Sibilant Affricate					tʃ dʒ						
	Fricative	ɸ β	f v	θ ð				ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
	Sibilant Fricative				s z	ʃ ʒ	ʂ ʐ					
	Lateral Fricative				ɬ ɮ							
	Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
	Tap / Flap		ⱱ		ɾ		ɽ					
	Trill	ʙ			r					ʀ		
	Approximant (Glide)		ʋ		ɹ		ɻ	j	ɰ			
	Co-articulated Approximant	ʍ w							ʍ w			
	Lateral Approximant (Liquid)				l		ɭ	ʎ	ɬ			

ARPA Consonants (pulmonic)		Place										
		Labial		Coronal					Dorsal			Laryngeal	
		Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Manner	Plosive / Stop	P B			T D				K G			
	Sibilant Affricate					CH JH						
	Fricative		F V	TH DH								HH
	Sibilant Fricative				S Z	SH ZH						
	Nasal	M			N				NG			
	Tap / Flap											
	Trill											
	Approximant (Glide)	W			R			Y	W			
	Lateral Approximant (Liquid)								L			

Figure 1
Comparison of the IPA consonant inventory and the derived ARPAbet articulatory inventory used in this work. The ARPAbet inventory was constructed from the L2-ARCTIC ARPAbet–IPA phoneme correspondence and organized according to IPA articulatory feature categories.

encode the articulatory structure underlying speech production. This is exemplified by L2-ARCTIC's selective phoneme inventory, further reducing phonetic granularity from 89 standard IPA phonemes excluding diacritics (International Phonetic Association 1999) to a reduced set of 40 phonemes. Notably, there is an absence of transcribed non-English articulations in the current constrained ARPAbet space despite being present in the corpus' L1 backgrounds. As a result, the resulting phonemic representations fail to capture the full richness and cross-linguistic variation essential for modeling accented or pathological speech.

To obtain articulatory features, each ARPAbet phoneme was mapped to a set of linguistically motivated articulatory attributes. IPA representations' articulatory mappings were used an as intermediate resource to define these. By decomposing ARPAbet phonemes into articulatory properties, the resulting feature representation allows the model to detect the corresponding ARPAbet phoneme while also being able to predict the particular features associated with it. Figure 1 illustrates how the reduced ARPAbet inventory collapses articulatory distinctions, leading to ambiguity and motivating the use of auxiliary articulatory supervision. The modeling process is further complicated by the existence of diphthongs and affricates. ARPAbet encodes these as single symbolic units, whereas IPA explicitly encodes their composite structure. This obscures internal phonetic structure and limits the model's ability to learn fine-grained temporal and articulatory transitions. In addition to this, phonemic transcriptions for both ARPA and IPA omit phonetic detail conveyed through diacritics and suprasegmental markers. We exhaustively enumerate these, as well as the absent non-native transcriptions, in the Appendix (Table A.1, Table A.2).

Taken together, these limitations highlight a critical gap between the phonetic richness required for accurate pronunciation modeling and the constrained symbolic representations used in current speech datasets. To address these limitations, we derive articulatory feature labels from ARPAbet phonemes, enabling the model to learn linguistically meaningful intermediate representations while maintaining compatibility with the corpus's phoneme inventory.

3.2 Articulatory Feature Multi-Task Learning Architecture

We introduce a hierarchical multi-task learning (HMTL) architecture for frame-level recognition from raw speech (Figure 2). The model decomposes phoneme prediction into several stages which can be grouped into first articulatory feature estimation, then phoneme classification, which is strengthened by the articulatory features. The architecture given an input waveform produces predictions for auxiliary articulatory features such as phoneme type, place, manner, voicing, height, backness, and roundedness to use those logits to strengthen a phoneme classification model.

Stage 1: Acoustic Feature Extraction with Learnable Shared Projections

We use a pretrained WavLM (Chen et al. 2022) model as the acoustic backbone to extract frame-level representation from raw waveform inputs. The model outputs hidden states from all of the transformer layers. All backbone parameters are initially frozen, allowing the auxiliary heads to learn an initial state of information utilizing a higher learning rate. Instead of relying on a single hidden layer from WavLM, we learn two combinations of layer representations. The acoustic representation is used primarily for articulatory prediction tasks and the phoneme representation is used for the final phoneme classification. Both representations are computed as a learnable weighted sum of hidden states h over L total WavLM layers:

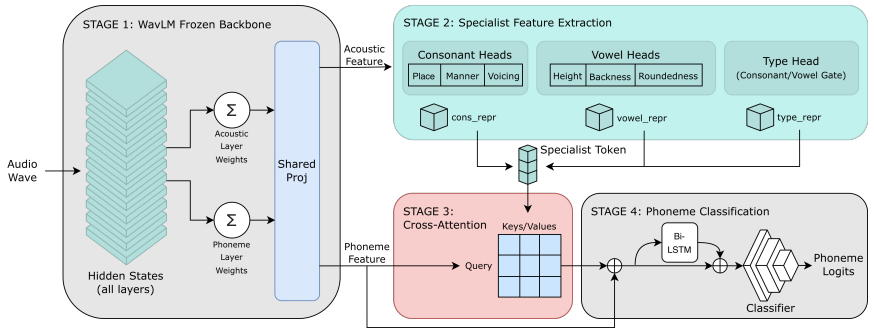


Figure 2
Hierarchical MTL architecture. (Stage 1) Input audio is first passed into a frozen WavLM backbone, where acoustic and phoneme representations are learned as two separate weighted sums of transformer layers. (Stage 2) Acoustic features are passed into separate articulatory feature estimation heads for consonant, vowel, and type representations. (Stage 3) These, along with the phoneme WavLM features are passed into a cross-attention layer, whose output is residually fused with the phoneme features. (Stage 4) This is then passed through a bidirectional LSTM to capture temporal dependencies before the final phoneme classification head.

$$\tilde{h} = \sum_{l=1}^{L} \alpha_l h^{(l)}, \quad \alpha = \text{softmax}(w)$$

where $w \in \mathbb{R}^L$ is a learnable parameter vector and α defines normalized layer weights. Both the acoustic and phoneme representations are then passed through a shared projection block, consisting of a linear transformation, layer normalization, GELU activation function, and dropout layer for training regularization.

Stage 2: Articulatory Specialist Heads

To model the phonological structure, the learned acoustic representation is passed into three articulatory specialist branches: a **phoneme type** branch, a **consonant** branch, and a **vowel** branch. These branches produce task-specific latent representations that are later used to guide phoneme classification. Each prediction head in the model follows a two-layer feedforward architecture. Specifically, each head consists of a linear projection from the input dimension to a hidden dimension, followed by a layer normalization, GELU activation, a dropout layer, and then a final linear layer that maps to the task output. This design provides a lightweight but expressive prediction module for each auxiliary task. The type branch is a binary classification head which predicts whether each frame corresponds to a consonant or vowel. This representation serves as a gating mechanism, allowing the model to distinguish between a vowel and consonant to strengthen the processing of the consonant/vowel articulatory heads.

Together, these three branches form a set of articulatory features that decompose speech into meaningful components before phoneme classification. By modeling articulatory structure, the model provides intermediate representations that can be leveraged in the downstream phoneme prediction stage.

Stage 3: Specialist Memory Cross-Attention To integrate articulatory knowledge into the phoneme prediction, we introduce a cross-attention mechanism that enables phoneme branch to dynamically attend to the specialized articulatory representations at each time step. This is performed by first introducing temporal concatenation as a structure in the architecture. Let $\mathbf{h}_t^{(p)} \in \mathbb{R}^d$ denote the phoneme-branch representation at frame t, and let $\mathbf{h}_t^{(c)}$, $\mathbf{h}_t^{(v)}$, $\mathbf{h}_t^{(y)} \in \mathbb{R}^d$ denote the consonant, vowel, and type specialist representations, respectively. These are stacked into a specialist memory matrix:

$$\mathbf{S}_t = \begin{bmatrix} \mathbf{h}_t^{(c)} \\ \mathbf{h}_t^{(v)} \\ \mathbf{h}_t^{(y)} \end{bmatrix} \in \mathbb{R}^{3\times d}.$$

The phoneme representation $\mathbf{h}_t^{(p)}$ serves as the query, representing the model's current estimate of phoneme-relevant acoustic context before incorporating articulatory information. The specialist memory is projected into key and value spaces via learnable matrices $W_K, W_V \in \mathbb{R}^{d\times d_k}$:

$$Q_t = \mathbf{h}_t^{(p)}, \quad K_t = \mathbf{S}_t W_K, \quad V_t = \mathbf{S}_t W_V, \quad K_t, V_t \in \mathbb{R}^{3\times d_k}.$$

The cross-attention output is then computed as:

$$\mathbf{c}_t = \mathrm{softmax}\left(\frac{\mathbf{h}_t^{(p)} K_t^\top}{\sqrt{d_k}}\right) V_t,$$

where the softmax produces normalized attention weights $\alpha_{t,i}$ over the three specialist tokens, and $\mathbf{c}_t = \sum_{i=1}^{3} \alpha_{t,i}\, \mathbf{v}_{t,i}$ is the resulting articulatory context vector. The scaling factor $1/\sqrt{d_k}$ stabilizes training by controlling the magnitude of the dot-product scores.

In contrast to simple concatenation, which treats all articulatory feature representations as equally informative, this cross-attention mechanism allows the learned phoneme representation to selectively attend to feature-specific signals at each time step. This is particularly important for resolving phoneme ambiguity, where the importance of articulatory dimensions (e.g., voicing, place, or manner) varies depending on the acoustic context. By dynamically weighting these feature representations, the model can more effectively disambiguate phonemes that are acoustically similar but differ along specific articulatory dimensions.

Stage 4: Fusion and Output branch

The articulatory context vector is fused with the phoneme representation via a residual connection:

$$\tilde{\mathbf{h}}_t = \mathrm{LayerNorm}\Big(\mathbf{h}_t^{(p)} + \mathbf{c}_t\Big).$$

The residual formulation preserves the original phoneme encoding while augmenting it with articulatory evidence. Layer normalization stabilizes training by reconciling the feature distributions of the two components, which originate from different latent subspaces. The fused representation $\tilde{\mathbf{h}}_t$ is passed through a bidirectional LSTM (Graves

and Schmidhuber 2005) to capture temporal dependencies:

$$\mathbf{h}_t^{\text{temporal}} = \text{BiLSTM}(\tilde{\mathbf{h}}_t) = \left[\overrightarrow{\mathbf{h}}_t \,;\, \overleftarrow{\mathbf{h}}_t\right],$$

where $\overrightarrow{\mathbf{h}}_t$ and $\overleftarrow{\mathbf{h}}_t$ encode past and future context, respectively. Because phonemes are influenced by their neighbors, frame-level predictions based on local features alone are often insufficient. The BiLSTM models contextual dependencies across time, smoothing noisy frame-level representations and propagating the attended articulatory features across the sequence. Finally, the temporal representation is passed to a final classification head:

$$\mathbf{y}_t = f_{\text{phoneme}}\left(\mathbf{h}_t^{\text{temporal}}\right),$$

producing frame-level phoneme logits for Connectionist Temporal Classification (CTC) decoding (Graves et al. 2006).

Multi-Task Learning Objective To jointly optimize phoneme recognition and articulatory feature prediction, we adopt a MTL framework with adaptive task weighting. Rather than relying on fixed loss weights, we employ a Pareto-based weighting that dynamically balances task objectives during training (Sener and Koltun 2018).

The primary phoneme task is trained with CTC loss:

$$\mathcal{L}_{\text{phoneme}} = \text{CTC}\left(\mathbf{y}^{(\text{phoneme})}, \hat{\mathbf{y}}^{(\text{phoneme})}\right),$$

where $\hat{\mathbf{y}}^{(\text{phoneme})}$ denotes the frame-level phoneme logits and $\mathbf{y}^{(\text{phoneme})}$ the ground-truth phoneme sequence. The auxiliary articulatory tasks use frame-level cross-entropy for multi-class targets (type, place, manner, height, and backness) and binary cross-entropy for binary targets (voicing and roundedness), each computed only over frames belonging to the relevant phoneme category (consonant or vowel).

The phoneme loss is assigned a fixed weight of 1.0, while the auxiliary task weights are determined via Multiple-Gradient Descent Algorithm (MGDA) (Désidéri 2012). Let $\mathbf{g}_i = \nabla_{\boldsymbol{\theta}_s} \mathcal{L}_i$ denote the gradient of the i-th auxiliary loss with respect to the shared representation $\boldsymbol{\theta}_s$. MGDA finds Pareto-optimal weights $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ by solving:

$$\min_{\boldsymbol{\alpha} \in \Delta^K} \left\| \sum_{i=1}^{K} \alpha_i \, \hat{\mathbf{g}}_i \right\|^2,$$

where Δ^K is the probability simplex and $\hat{\mathbf{g}}_i = \mathbf{g}_i / \|\mathbf{g}_i\|$ are the ℓ_2-normalized task gradients. This quadratic program is solved efficiently via the Frank–Wolfe algorithm at each training step. The overall training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{phoneme}} + \lambda_s \sum_{i=1}^{K} \alpha_i \, \mathcal{L}_i,$$

where λ_s is dependent on the stage and is adjusted throughout the training phases to progressively focus on the primary phoneme task. By decoupling the phoneme loss from the Pareto optimization, we ensure that auxiliary articulatory tasks inform the

shared representation without harming phoneme convergence. The MGDA weights prevent any single auxiliary task from dominating the gradient, promoting balanced learning across the articulatory feature space.

Training Strategy

To ensure stable optimization and effective learning of both acoustic and articulatory representations, we adopt a progressive training strategy consisting of two stages. In the first stage, the pretrained WavLM backbone is fully frozen, and only the newly introduced components are trained. During this phase, we use a relatively higher learning rate for the trainable parameters. This allows the newly initialized layers to adapt to downstream phoneme recognition tasks and learn meaningful representations without interfering with pretrained acoustic features. Since the WavLM representation starts off as fixed, the heads learn a stable mapping without gradient interference. In the second stage, we unfreeze the top k layers of the WavLM encoder and jointly optimize both the backbone and head parameters. Unfreezing only the higher layers of WavLM allows the model to refine higher-level acoustic representations while preserving low-level features during the pretraining. To prevent catastrophic forgetting and overfitting, we adopt a smaller learning rate for both the backbone and head parameters. In addition, we incorporate a replay mechanism in which a subset of previously seen samples are periodically revisited during training to further mitigate catastrophic forgetting and stabilize the fine-tuning process. Overall, the training procedure can be interpreted as a two-phase optimization problem: an initial representation alignment phase, followed by a joint refinement phase. The first phase learns a stable mapping from pretrained features, while the second phase refines both the head and higher layer WavLM features jointly.

3.3 Momentum-Based Pseudo Labeling

As discussed previously, phoneme labels are often noisy and limited. These challenges are further intensified in pathological and non-canonical speech, where rarer conditions may have little to no labeled data available. To improve generalization in such settings and leverage additional unlabeled speech, we extend our multi-task framework with a semi-supervised momentum-based learning strategy called momentum pseudo labeling. This approach integrates labeled data with a teacher-student distillation setup applied to unlabeled inputs, enabling the model to learn more robust representations.

Data Setup

1. Labeled dataset $\mathcal{D}_L = \{(x_i, y_i)\}$: Provides phoneme sequences and articulatory feature annotations for all tasks.

2. Unlabeled dataset $\mathcal{D}_U = \{x_j\}$: Contains raw speech waveforms without annotations, used for consistency-based learning.

For unlabeled samples, a clean waveform x is used by the teacher model and an augmented waveform x is used by the student model.

Momentum Teacher

We maintain a teacher model $f_{\theta'}$ whose parameters are updated as an exponential moving average (EMA) of the student model f_θ:

$$\theta' \leftarrow \alpha\theta' + (1 - \alpha)\theta$$

where $\alpha \in [0, 1)$ controls the update rate.

This produces a temporally smoothed model that serves as a source of pseudo-labels for unlabeled data.

Learning on Unlabeled Data

For unlabeled inputs $x \sim \mathcal{D}_U$, the teacher produces soft pseudo-labels from the clean signal:

$$p_{\theta'}(y \mid x)$$

The student processes an augmented version x^{aug} and produces:

$$p_{\theta}(y \mid x^{\text{aug}})$$

Rather than using strict pseudo-labels, we enforce consistency between the two distributions:

$$\mathcal{L}_{\text{unsup}} = \mathbb{E}_{x \sim \mathcal{D}_U} \left[D \left(p_{\theta}(y \mid x^{\text{aug}}),\, p_{\theta'}(y \mid x) \right) \right]$$

The total unsupervised loss consists of a KL divergence for all of the categorical heads:

$$\mathcal{L}_{\text{cat}}^{\text{unsup}} = \text{KL} \left(p_{\theta'} p_{\theta} \right) \tag{1}$$

The L2 distance is used for the binary heads:

$$\mathcal{L}_{\text{bin}}^{\text{unsup}} = \sigma(z_{\theta'}) - \sigma(z_{\theta})^2 \tag{2}$$

The loss is then aggregated as a sum. The complete objective combines the total supervised loss and unsupervised loss.

Although some explicit frame-level annotations are unavailable, the teacher produces a sequence of probability distributions over phoneme tokens at each time step. These distributions are soft signals allowing the student model to learn where phonemes occur without requiring explicit labels. By minimizing the KL divergence between the teacher and student outputs, the model is able to produce more stable predictions even with noisy data.

3.4 Speech-based Data Augmentations

To improve robustness under limited supervision, we apply waveform-level speech augmentations that increase acoustic variability. For an input waveform x, augmentation is applied with probability p_{aug}:

$$\tilde{x} = \begin{cases} \mathcal{A}(x; \theta), & \text{with probability } p_{\text{aug}}, \\ x, & \text{otherwise}, \end{cases} \tag{3}$$

where $\mathcal{A}$ is one augmentation operator and θ denotes its sampled parameters. The available operations are phase perturbation (Lei et al. 2023), VTLP (Jaitly and Hinton 2013), pitch shifting, prosodic time-stretching, and optional additive noise. Their strengths are sampled from predefined ranges in the configuration. Time and frequency masking are

implicitly provided by the WavLM frontend through SpecAugment (Park et al. 2019), and are therefore not re-implemented.

Phase perturbation. Phase perturbation is performed in the STFT domain. Given

$$X(\omega, \tau) = |X(\omega, \tau)| e^{j\phi(\omega,\tau)}, \tag{4}$$

we preserve the magnitude and perturb the phase:

$$\tilde{X}(\omega, \tau) = |X(\omega, \tau)| e^{j(\phi(\omega,\tau)+\Delta\phi(\omega,\tau))}. \tag{5}$$

The waveform is then reconstructed by inverse STFT. Our code supports standard, frequency-dependent, temporally smoothed, and selective-band variants, with perturbation strength sampled from a configurable interval (default $[0.05, 0.25]$). This augmentation is intended to mimic unstable or effortful phonation, as observed in spastic dysarthria.

Vocal Tract Length Perturbation (VTLP). VTLP warps the frequency axis,

$$f' = g_\alpha(f), \tag{6}$$

where α is sampled from a configured warp range. In our implementation, VTLP is applied over the full signal with coverage 1.0 and cutoff frequency $f_{\text{hi}} = 4800$ Hz. This transformation increases variability in the spectral envelope and formant structure, which is relevant to resonance-related deviations such as breathy or dysphonic speech, while also improving invariance to speaker-dependent spectral variation.

Pitch and prosodic perturbation. Pitch shift modifies the perceived fundamental frequency by a sampled number of semitone steps n, while prosodic augmentation applies time-stretching with rate γ:

$$\tilde{x} = \mathcal{P}(x; n), \qquad \tilde{x}(t) = x(\gamma t). \tag{7}$$

The values of n and γ are sampled from configured ranges. These augmentations target abnormalities in intonation, speech rate, and stress patterning, and are motivated by disorders such as ataxic, hypokinetic, and hyperkinetic dysarthria, as well as CAS, where prosodic control is frequently disrupted.

Additive noise. When enabled, noise is added using an externally provided noise source with signal-to-noise ratio sampled from configured bounds. Although additive noise is not a physiological model of disordered speech, it can approximate degraded or breathy voice quality and improves robustness to low-quality acoustic realizations.

Overall, these augmentations should be interpreted as *acoustic proxies* rather than faithful clinical simulations. They do not reproduce the neuromotor mechanisms or coordinated error patterns of real pathological speech, but instead expand the training distribution along clinically relevant dimensions of spectral, phonatory, and prosodic variability.

4. Experimental Setup

4.1 Dataset description.

We train and evaluate on the L2-ARCTIC corpus, which contains read speech from 24 non-native English speakers spanning six first-language (L1) backgrounds: Vietnamese, Korean, Mandarin, Spanish, Hindi, and Arabic, with four speakers per L1 (balanced for gender)(Zhao et al. 2018). Each utterance is accompanied by Praat TextGrid (Boersma and Weenink 2001) annotations providing both *canonical* phoneme transcriptions (the expected pronunciation) and *perceived* phoneme transcriptions (what expert annotators actually heard), along with time-aligned segment boundaries. As previously stated and as supported by previous studies (Farish, Davis, and Wilson 2020; Flege 1995; Best and Tyler 2007), while this does not explicitly include pathological data, we use the L2 speech as a proxy for pathological speech, mainly noting similarities to dysarthritic speech.

Ground-Truth Establishment.

Defining ground-truth phoneme labels for non-native speech is inherently challenging: phoneme productions by L2 speakers often fall between canonical categories, and annotator judgments can vary considerably. Prior work on non-native speech transcription reports that inter-annotator agreement at the phone level typically reaches only fair-to-moderate levels, with Cohen's Kappa values often in the range of 0.7–0.85 (Ryu, Kim, and Chung 2012), depending on annotation constraints and task design. The L2-ARCTIC annotations were produced by trained linguists experienced in transcribing non-native speech, with automated consistency checks and secondary verification applied to improve quality. Nevertheless, the inherent subjectivity of labeling accented and mispronounced speech means that ground-truth labels are noisy by nature—a key motivation for our use of MPL, which can leverage unlabeled data to reduce dependence on potentially inconsistent annotations.

Preprocessing. We apply normalization to the raw annotations before training:

1. **Stress removal.** ARPAbet symbols include lexical stress markers (e.g., `AH0`, `AH1`, `AH2`). We strip all stress digits, collapsing stressed variants into a single phoneme identity (e.g., `ah`). This reduces the label space without discarding segmental information, since stress distinctions are not the focus of articulatory assessment.
2. **Silence normalization.** Silence tokens (`sil`, `sp`, `spn`, `pau`) are unified under a single `sil` label. Artificial silences introduced by the L2-ARCTIC annotation scheme are removed, and consecutive segments are collapsed into a single token to prevent the model from over-representing pauses.
3. **Phoneme inventory.** After normalization, the inventory comprises 39 non-silence phonemes plus a silence token. For CTC training, silence frames are excluded from the target sequence, yielding a 39-phoneme vocabulary with an appended CTC blank, for a total of 40 output classes.

Articulatory Feature Decomposition.

The seven auxiliary targets are derived deterministically from our ARPAbet-to-articulatory-feature mapping and are computed at the frame level using the time-aligned segment boundaries from the TextGrid annotations. Consonant-specific features

(place, manner, voicing) are masked on vowel frames and vice versa, so each auxiliary head is supervised only on phonologically relevant frames.

Speaker Splits. We partition speakers into 18 for training and 6 for testing, ensuring no speaker overlap between sets. The test speakers represent one speaker from each of the six L1 backgrounds. The training set is further split 90/10 by utterance for training and validation.

4.2 Experiments overview

Ablation Study

MTL Heads To systematically evaluate the contribution of each articulatory auxiliary task to phoneme recognition, we conduct a Leave-One-Out (LOO) ablation study within the proposed multi-task learning (MTL) framework discussed in section 3.2. This analysis is designed to isolate the effect of individual articulatory features.

In every iteration one articulatory feature head is dropped, following the flow:
Let $\mathcal{T}$ denote the set of active tasks and $f_\theta^{\mathcal{T}}$ the corresponding model.
The full task set is:

$$\mathcal{T}_{\text{full}} = \{\text{phoneme}, \text{type}, \text{place}, \text{manner}, \text{voicing}, \text{height}, \text{backness}, \text{roundedness}\}.$$

For each auxiliary task $t \in \mathcal{T}_{\text{aux}}$, we define an ablated model:

$$\mathcal{T}^{(-t)} = \mathcal{T}_{\text{full}} \setminus \{t\}, \quad f_\theta^{(-t)} = f_\theta^{\mathcal{T}^{(-t)}}.$$

Each model $f_\theta^{(-t)}$ is trained under identical conditions, and performance differences are used to quantify the contribution of task t.

Finally, rather than adopting a conventional soft ablation approach, where task losses are masked while retaining the full model architecture, we employ a hard ablation strategy that modifies the network structure. Specifically, when an articulatory task is removed, its corresponding prediction head is eliminated entirely from the model.

All ablation experiments are conducted under a simplified training regime, where data augmentations and momentum pseudo-labeling are disabled, and both Stage 1 (frozen backbone) and Stage 2 (progressive unfreezing) are limited to 5 epochs each. Although this setup does not fully converge the model, it provides a consistent basis for comparing architectural variants. As the ablation study focuses on relative differences between configurations rather than absolute performance, this regime is sufficient to capture the contribution of individual components.

Data Augmentation To assess the contribution of each waveform-level augmentation described in Section 3.4, we conduct a Leave-One-Out (LOO) ablation study over the augmentation pipeline. Starting from the full augmentation setting, which includes phase perturbation, VTLP, pitch shifting, prosodic time-stretching, and optional additive noise, we remove one augmentation at a time while keeping all others unchanged. Each model is trained under the same architecture, optimization settings, and data split, such that performance differences can be attributed to the removed augmentation. Unlike the MTL head ablation, this analysis does not modify the network structure, but only the stochastic training-time transformation pipeline. The purpose of this experiment is to quantify the extent to which each augmentation contributes to robustness by increasing acoustic variability along spectral, phonatory, and prosodic dimensions.

4.3 Implementation Details

The model was implemented in PyTorch with HuggingFace Transformers, using `microsoft/wavlm-base-plus` as the acoustic encoder. Unlike ASR-fine-tuned encoders, `wavlm-base-plus` is used here as a general-purpose pretrained speech encoder and does not encode an explicit audio-text mapping before task-specific fine-tuning. The Base-Plus variant was pretrained on 94k hours of 16 kHz speech drawn from Libri-Light, GigaSpeech, and English VoxPopuli, providing substantially broader acoustic coverage than standard base-scale SSL speech models trained on smaller corpora. Training was conducted on NVIDIA RTX A4000, RTX 3090, and GV100 GPUs.

We used a cascaded multi-task architecture with hidden dimension 384 and dropout 0.35. Training proceeded in two stages. In Stage 1, the WavLM backbone was frozen and only the task-specific layers were trained for 15 epochs. In Stage 2, the last 6 WavLM transformer layers were unfrozen and jointly optimized with the task heads for 15 additional epochs. Full-backbone fine-tuning was not used in the final setting.

Optimization used AdamW with separate parameter groups for the task heads and WavLM backbone. The learning rate was 2.00×10^{-3} in Stage 1 for the task heads, and 9.78×10^{-5} / 9.58×10^{-5} for the task heads / WavLM backbone in Stage 2. A cosine annealing scheduler was applied in Stage 2, and gradient clipping with maximum norm 0.35 was used throughout. We trained the model with an MGDA-based multi-

Table 1
Model performance on L2-ARCTIC (Phoneme Error Rate, %)

Model	PER ↓	Prec	Rec	F1
Canonical Baselines				
Wav2Vec2Phoneme	22.76	0.311	0.368	0.337
WavLM-base-plus	28.02	0.281	0.477	0.353
Fine-Tuned Baselines (WavLM-base-plus)				
Wav2Vec2Phoneme	14.81	0.540	0.479	0.507
WavLM	14.71	0.545	0.508	0.526
WavLM + Aug	14.72	0.539	0.467	0.500
WavLM + MPL	14.54	0.551	0.483	0.515
WavLM + Aug + MPL	14.46	0.548	0.462	0.502
Multi-Task Learning (WavLM-base-plus backbone)				
Parallel MTL	14.82	0.532	0.449	0.487
Parallel MTL + Aug.	14.55	0.544	0.462	0.500
Parallel MTL + Aug. + MPL	14.46	0.546	0.459	0.499
Hierarchical MTL (HMTL)	14.23	0.554	0.534	0.544
HMTL + Aug.	13.89	0.570	0.488	0.526
HMTL + Aug. + MPL	13.68	0.576	0.514	0.543
HMTL + Cross-attention (CA)	13.60	0.584	0.519	0.549
HMTL + CA + Aug.	13.52	0.587	0.511	0.546
HMTL + CA + Aug. + MPL	**13.47**	0.586	0.524	0.553
External Baselines				
$MV_{multi} - MT_{seq}$ (EL Kheir, Chowdhury, and Ali 2023)	14.13	0.614	0.592	0.603

task objective and CTC phoneme supervision (blank index 40). PER was computed by greedy CTC decoding followed by Levenshtein distance against the reference phoneme sequence, and validation PER was used for model selection. Hyperparameters were tuned with Optuna using a TPE sampler and MedianPruner. The search included hidden dimension, dropout, batch size, stage-wise learning rates, stage lengths, and the number of unfrozen WavLM layers. The final configuration was selected by minimizing validation PER.

5. Results and Discussion

5.1 Model Evaluation and Comparisons

We evaluate multiple model variants: baseline phoneme recognition models such as Wav2Vec and WavLM, augmented baselines, and our proposed multi-task learning (MTL) architectures (Table 1), which outperforms the multi-view sequential MTL framework proposed by EL Kheir, Chowdhury, and Ali (2023). Performance is measured using phoneme error rate (PER). We see that zero-shot performance (where we do not train on a subset of the L2-Arctic dataset) is significantly lower than all of the non-canonical fine-tuned models. This addresses our first research question and paints a strong motivation for fine-tuning on a non-canonical speech dataset. Among fine-tuned baseline systems, WavLM slightly outperforms Wav2Vec2, likely due to its utterance mixing strategy and gated relative positional bias, which better capture temporal dependencies in speech. Incorporating data augmentation consistently improves performance, suggesting that exposure to greater acoustic variability improves model robustness, particularly for unstable or low-energy phonetic segments. Incorporating MPL yields further gains across model variants. While MPL consistently improves overall PER, its effect on individual error types depends on the underlying architecture. As shown in Table 2, MPL reduces insertion and substitution errors for both the baseline and HMTL model (-10.8% and -11.8%, respectively), although this is accompanied by increased deletions ($+4.8\%$ and $+16.9\%$), suggesting a more conservative decoding strategy. In contrast, when cross-attention is introduced into HMTL, MPL instead reduces deletions (-9.7%) while slightly increasing insertions and substitutions. Despite these differing

Table 2
Changes in phoneme error types after applying MPL.

Model	Error Type	Before	With MPL	Δ (%)
Baseline	Substitutions	3055	3006	-1.60
	Deletions	673	705	4.76
	Insertions	555	495	-10.81
	Cross-class	123	117	-4.88
HMTL	Substitutions	2860	2793	-2.34
	Deletions	457	534	16.85
	Insertions	542	478	-11.8
	Cross-class	179	173	-3.35
HMTL with cross-attention	Substitutions	2699	2732	1.22
	Deletions	576	520	-9.72
	Insertions	488	501	2.66
	Cross-class	169	165	-2.37

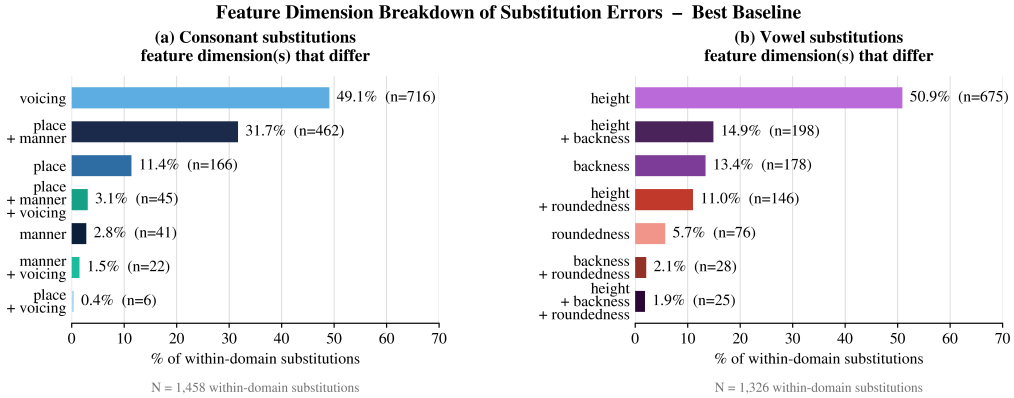


Figure 3
Breakdown of within-domain substitution errors by articulatory feature dimensions for the best baseline model.

error profiles, overall PER improves, indicating that MPL primarily enhances generalization and sequence alignment, with its influence on insertion-deletion trade-offs depending on the underlying architecture. Introducing MTL in a parallel formulation yields marginal improvements over the baseline, suggesting that naively incorporating auxiliary objectives does not improve performance and may instead introduce optimization challenges. A more substantial improvement is observed with hierarchical MTL architectures and cross-attention, which explicitly structure the learning of articulatory features by progressively supervising higher-level representations. These results indicate that performance gains are not driven by any single component, but by the interaction between model architecture, linguistic structure, data augmentation, and semi-supervised learning. While augmentation and MPL enhance generalization, the largest improvements arise when phoneme prediction is built with linguistically motivated auxiliary tasks. This motivates the subsequent analysis of error patterns to better understand where these improvements originate.

Best Baseline Model Error Analysis. We first decompose same-class substitution errors by articulatory feature differences for the strongest baseline model (PER = 14.46%) 3). We analyze substitution errors since they constitue the majority of errors, indicating that model performance is primarily limited by confusion between acoustically similar phonemes rather than segmentation failures. The comprehensive breakdown is outlined in Appendix (Tables B.2 and B.3). For same-class substitutions, the errors exhibit highly structured patterns and predominantly involve low-dimensional feature changes. For consonants, nearly half of all substitutions differ only in voicing (49.1%), while a further 31.7% differ jointly in *place* and *manner*. Isolated place and manner errors are comparatively uncommon, indicating that consonant confusions are dominated by a small number of articulatory dimensions. Similarly, vowel substitutions are driven by *height* differences (50.9%), followed by combined *height-backness* deviations. Overall, substitutions are concentrated along specific articulatory dimensions, with most confusions arising from differences in a single feature. This distribution is consistent with established phonetic properties of speech. Voicing is encoded through low-frequency periodicity that can be masked by noise, while vowel height is encoded through varying

Table 3
Changes in recognition errors and articulatory feature substitutions between the best baseline (WavLM + Aug + MPL) and the best proposed model (HMTL + CA + Aug + MPL). Δ Domain Share is computed within consonant or vowel substitution classes and denotes normalized shift in error distribution.

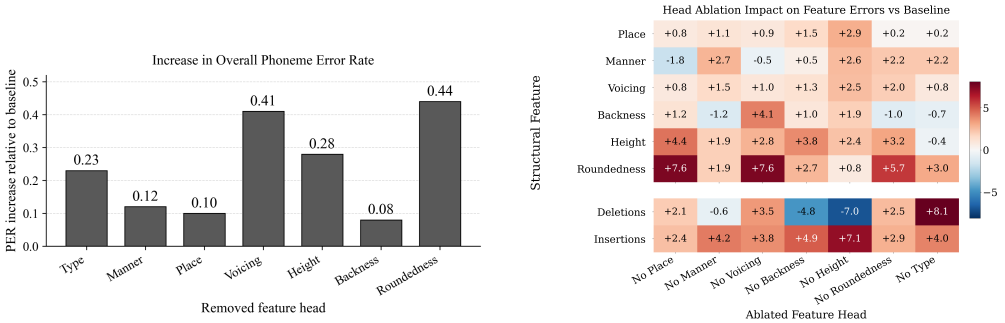
Category	Metric	Δ (%)	Δ Domain Share (%)
Type	Cross-class	$+41.0$	-
Articulatory	Place	-3.24	$+1.01$
	Manner	-3.86	$+0.66$
	Voicing	-10.14	-1.67
	Height	-5.65	-0.06
	Backness	-5.83	-0.07
	Roundedness	-4.73	$+0.14$

formant structure that can overlap across categories. Consequently, the baseline model captures coarse phonetic categories but fails to resolve fine-grained distinctions within them. These observations motivate our proposed MTL framework, which improves phoneme discrimination through articulatory supervision.

To assess the impact of MTL, we examine substitution errors across articulatory feature dimensions (Table 3). Relative to the strongest baselines, the proposed model reduces substitution errors across every feature, indicating more accurate articulatory discrimination under MTL. Reductions are concentrated along the most error-prone features, voicing and vowel height, which experience the some of the largest reductions (-10.1% and -5.65%), respectively, whereas place and manner exhibit the smallest gains (-3.24% and -3.86%). When normalized within each domain, voicing, the dominant source of consonant errors, decreases more rapidly than the remaining features, resulting in a redistribution of residual errors to place and manner. In contrast, in vowels, the relative distributions changes only marginally. Notably, we observe an increase in cross-class confusions under MTL. This can be be attributed to the lack of enforcement between the consonant-vowel boundary as within-class ambiguities are reduced, therefore increasing errors between /R/ and /ER/. These findings support the hypothesis that explicitly modeling articulatory features enables the model to disentangle correlated dimensions, thereby resolving the structural confusions observed in the baseline.

5.2 Multi-Task-Learning Ablations

The results of the articulatory task head ablation study are summarized in Figure 4. The configuration with all heads achieved the lowest phoneme error rate (PER) of 14.13%. Removing any individual articulatory head consistently degraded performance, with PER increases ranging from +0.08 to +0.44 relative to the baseline. The most substantial degradations were observed when removing the Roundedness (+0.44) and Voicing (+0.41) heads, followed by Height (+0.28) and Type (+0.23). In contrast, removing Backness (+0.08), Place (+0.10), and Manner (+0.12) resulted in comparatively smaller performance drops. These results indicate that not all articulatory features contribute equally, and that certain features, particularly Roundedness and Voicing, encode highly informative signals that are critical for accurate phoneme recognition.



(a) Aggregate PER increase by ablated head. **(b)** Head-level PER deltas across features.

Figure 4

Leave-one-out attention head ablation analysis of phoneme error rate (PER). (a) Aggregate increase in PER when ablating heads associated with each phonological feature, showing the relative contribution of different task heads. (b) Heatmap of PER changes for individual head removals (x-axis) across articulatory feature error types (y-axis), where color encodes the change in error relative to the baseline model.

A more fine-grained view of these effects is shown in the error-type breakdown heatmap (Figure 4b. Removing a feature head generally increases errors associated with that feature, demonstrably, removing roundedness significantly increases roundedness-related errors (+5.7%) while also amplifying insertion (+2.9%) and deletion errors (+2.5%). These increased segmentation errors suggest that this feature plays a key role in stabilizing vowel boundaries. However, the effects extends beyond single features, revealing inter-dependencies between articulatory dimensions. Namely, removing the Voicing head leads to increases not only in voicing, but also in roundedness (+7.6%), backness (+4.1%), and insertions (+3.8%), indicating cross-feature dependencies. In contrast, some removals lead to partial compensatory effects. To illustrate, removing Backness reduces deletion errors (-4.8%) but increases insertion errors (+4.9%), highlighting trade-offs in how the model distributes phonological information.

From a linguistic perspective, features such as voicing rely on low-frequency periodicity that can easily be masked by noise (Stevens 2000; Ladefoged and Johnson 2014), making them inherently difficult to model, but highly valuable when learned, thereby creating substantial errors when removed. Height and Backness ablations may cause deletions to decrease as both are continuous, within-vowel features encoded through formant structure (F1 and F2), which varies smoothly across vowel realizations and provide weaker discriminative boundaries for phoneme presence detection. Removing these features reduces sensitivity to fine-grained vowel distinction, making the model less likely to reject ambiguous vowel segments and more prone to insertion and substitution errors. A similar logic applies to other localized blue cells: removing the Place and Voicing head slightly reduces manner-related errors, but omitting Manner yields small improvements in backness and deletions. Notably, some ablations lead to localized reductions in specific error types (blue regions), which may appear counterintuitive. For instance, removing Place or Voicing reduces manner related errors, while removing the Manner head improves backness and deletion errors. Place and manner are supported by partially redundant acoustic cues, including spectral shape, temporal structure, and formant transitions (Stevens 2000; Ladefoged and Johnson 2014), making these more

robust to the removal of any single supervisory signal, resulting in even improved errors under ablation. However, these effects do not necessarily reflect direct linguistic similarity between features, but rather interactions within the learned representation that arise from ablations. Jointly learning features with competing representations can introduce representational competition, leading to localized improvements in some features when ablated, despite overall degradation in performance. The fact that removing any head worsens PER, while simultaneously redistributing feature-specific errors in structured ways, supports the claim that MTL improves performance not by uniform improvements, but by disentangling correlated articulatory dimensions according to linguistic reasoning.

5.3 Augmentation Ablations

Table 4
Leave-one-out ablation of audio augmentations.

Configuration	PER (%)
All augmentations	**13.49**
− Noise	13.76
− VTLP	13.80
− Phase perturbation	13.83
− Prosody (Time-stretch)	13.87
− Pitch shift	13.94

To understand the contribution of different data augmentation strategies on model robustness, we performed a leave-one-out ablation study. We started with a baseline utilizing all augmentations (pitch shifting, time-stretching for prosody, phase perturbation, VTLP, and additive noise) and systematically removed one at a time. Overall, using our best-performing model configuration of HMTL+c.a.+MPL, our full augmentation suite achieves the lowest PER of 13.49% (Table 4), demonstrating that a diverse set of regularization is strictly beneficial.

Pitch and Prosody.

Removing pitch shifting causes the most severe performance degradation, increasing PER to 13.94%. This aligns closely with the phonetic characteristics of L2 speech: non-native learners exhibit substantial fundamental frequency (F_0) variance, often inappropriately transferring L1 intonation or tone patterns to the L2 (Mennen 2015). Because phonemic contrasts in English rely heavily on the spectral envelope (formants) rather than absolute pitch, forcing the model to be invariant to F_0 prevents it from overfitting to irrelevant pitch contours. Similarly, removing prosodic time-stretching yields the second highest error (13.87%). L2 learners are characterized by highly variable speaking rates, frequent hesitations, and inconsistent vowel elongation. Time-stretching naturally mimics these temporal distortions, improving the model's ability to locate phoneme boundaries despite dysfluent rhythms.

Vocal Tract and Phase

VTLP and phase perturbation address structural acoustic variations rather than behavioral ones. VTLP simulates speakers with differing physical vocal tract lengths, while phase perturbation simulates variations in microphone placement and room acoustics without fundamentally altering the linguistic magnitude spectrum. While less

impactful than pitch and prosody, removing phase perturbation ($+0.34\%$ PER) and VTLP ($+0.31\%$ PER) still noticeably degrades performance, confirming that simulating diverse physical profiles and recording environments aids generalization safely.

Naturalness

Crucially, these augmentations mimic naturalistic variations better than purely synthetic perturbations (like masking or dropout). By restricting pitch shifts and time stretches within physically plausible bounds, we systematically expose the network to the specific axes of variance, pitch instability and temporal hesitation, that define the pathological and L2 speech distribution.

6. Limitations

Our approach introduces a structured framework for phoneme recognition using articulatory feature decomposition, but several limitations remain. The model operates at the phonemic level using a reduced ARPAbet inventory, which collapses fine-grained phonetic variation include allophonic distinctions such as aspiration, flapping, and vowel reduction, limiting the system's ability to capture subtle but clinically relevant pronunciation errors. This is further compounded by ARPAbet itself, which omits several articulatory distinctions present in IPA, reducing the granularity of both supervision and evaluation. Although some phonemes posses multiple articulatory properties simultaneously, each auxiliary head is trained to predict a single discrete label. For example, /W/ exhibits both labial and dorsal articulations, yet the place classifier can predict only a single place category. This limits the model's ability to capture complex articulatory structure. Furthermore, the proposed framework focuses exclusively on segmental phoneme recognition and does not model suprasegmental aspects of speech, including stress, rhythm, and intonation. As another limitation articulatory features are modeled as discrete categorical targets rather than continuous physiological processes. The framework therefore does not explicitly capture articulatory dynamics such as airflow, timing, or motion. Future work could explore richer inventories such as IPA that more faithfully represent these dynamics. Such extensions would allow the framework to better capture co-articulation, and the inherently continuous nature of speech production. Finally, an additional limitation of our work is that we evaluate only on L2-Arctic rather than a pathological dataset. We do this out of concerns for data pollution regarding existing large-scale pre-trained backbones being trained on most labeled phoneme and transcribed audio datasets, including older pathological speech datasets. L2-Arctic is relatively recent and has not, to our knowledge, been utilized as training data for the baselines we have chosen in this study. We thus motivate the research community for the creation of newer phoneme-labeled pathological speech datasets.

7. Conclusion

Pathological phoneme recognition remains challenging due to structured deviations in speech, limited data availability, and noisy or ambiguous labels. Conventional models, which treat phonemes as independent categories, are not well-equipped to capture these structured variations, particularly under data-constrained settings. In this work, we introduce an articulatory-feature-based architecture for phoneme recognition that explicitly incorporates physiological structure, and evaluate on L2 speech as a controlled proxy for systematic non-canonical and pathological phoneme variation. By

modeling phoneme recognition as a composition of articulatory feature prediction tasks, integrating these representations through a cross-attention mechanism, and projecting into phoneme space after temporal modeling, our approach addresses structured errors that arise from independent phoneme modeling in standard baselines.

We demonstrate these improvements empirically through consistent reductions in phoneme error rate across strong baselines and ablated variants. Most noticeably the largest accuracy improvement was seen when combining our multi-task-learning framework with pseudo-labeling and strong augmentations, improving accuracy significantly compared to baseline methods even with the same data diversity techniques applied. This suggests that sufficient data diversity is necessary to fully realize the benefits of structured phoneme modeling.

In addition, we observe systematic reductions in structured error patterns. This is shown by an absolute decrease in errors and a redistribution away from within-feature confusions toward cross-feature errors. When analyzing error reduction across features, the addition of MTL affects the particular features that contributed most to errors observed in the baseline. Ablation studies further show that removing individual articulatory feature heads leads to predictable, linguistically interpretable degradation in performance. Vowel features in particular prove to be highly impactful; we find that including the vowel roundedness task head significantly reduces associated roundedness recognition error rates, and that the consonant voicing head further helps this category of confusions. While voicing and roundedness are distinct features, they may interact indirectly in the acoustic signal through co-articulation or overlap and other learned statistical regularities. We therefore interpret their joint importance in the model as reflecting correlated acoustic cues rather than a direct linguistic dependency.

Overall, our results highlight the importance of incorporating articulatory structure into speech recognition systems, particularly for non-canonical and clinical speech settings. While our experiments use accented speech as a proxy, future work should extend this framework to clinically validated pathological datasets and explore richer articulatory representations, including multi-label and continuous formulations. Improving pathological phoneme recognition has the potential to support digital clinical tools and enhance feedback in speech-language pathology workflows, motivating future attention in this field.

8. References

References

American Speech-Language-Hearing Association. 2023. Poll shows increases in hearing, speech, and language referrals, more communication challenges in young children.

Baevski, Alexei, 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.

Berisha, Visar and Julie M. Liss. 2024. Responsible development of clinical speech ai: Bridging the gap between clinical research and technology. *npj Digital Medicine*, 7:208.

Best, Catherine T. and Michael D. Tyler. 2007. Nonnative and second-language speech perception: Commonalities and complementarities. In Murray J. Munro and Öcke-Schwen Bohn, editors, *Second language speech learning: The role of language experience in speech perception and production*. John Benjamins, Amsterdam, pages 13–34.

Boersma, Paul and David Weenink. 2001. Praat, a system for doing phonetics by computer. *Glot international*, 5:341–345.

Carnegie Mellon University. 1998. The cmu pronouncing dictionary. `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`. Defines ARPAbet phoneme set.

Chen, Sanyuan, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 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.

Cleveland Clinic. 2025. Dysarthria (slurred speech): Symptoms, causes & treatment. Medically reviewed; Accessed: 2026-04-11.

Cohen, Ariel, Denis Shyrman, Aleksandr Solonskyi, Roman Frenkel, Arkady Krishtul, and Oren Gal. 2025. Robust prosody modeling for synthetic speech detection. *Speech Communication*, 174:103283.

Duffy, Joseph R. 2019. *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*, 4 edition. Mosby, St. Louis, MO.

Désidéri, Jean-Antoine. 2012. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5):313–318.

EL Kheir, Yassine, Shammur Chowdhury, and Ahmed Ali. 2023. Multi-View Multi-Task Representation Learning for Mispronunciation Detection. In *9th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 86–90.

Farish, Brian A., Lori A. Davis, and Laura D. Wilson. 2020. Listener perceptions of foreignness, precision, and accent attribution in a case of foreign accent syndrome. *Journal of Neurolinguistics*, 55:100910.

Fatehi, Kavan, Mercedes Torres Torres, and Ayse Kucukyilmaz. 2025. An overview of high-resource automatic speech recognition methods and their empirical evaluation in low-resource environments. *Speech Communication*, 167:103151.

Flege, James Emil. 1995. Second language speech learning: Theory, findings and problems. In Winifred Strange, editor, *Speech perception and linguistic experience: Issues in cross-language research*. York Press, Baltimore, pages 233–277.

Glocker, Kevin and Munir Georges. 2024. *Hierarchical Multi-task Learning with Articulatory Attributes for Cross-Lingual Phoneme Recognition*. Springer International Publishing, Cham.

Graves, Alex, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, Association for Computing Machinery, New York, NY, USA.

Graves, Alex and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610.

Hosom, John-Paul. 2009. Speaker-independent phoneme alignment using transition-dependent states. *Speech Communication*, 51(4):352–368.

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press.

Jaitly, Navdeep and Geoffrey E Hinton. 2013. Vocal tract length perturbation (vtlp) improves speech recognition. In *Proc. ICML workshop on deep learning for audio, speech and language*, volume 117, page 21.

Kent, Ray D and Y J Kim. 2003. Toward an acoustic typology of motor speech disorders. *Clin. Linguist. Phon.*, 17(6):427–445.

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

Ladefoged, Peter and Keith Johnson. 2014. *A Course in Phonetics*, 6 edition. Cengage Learning.

Lei, Chengxi, Satwinder Singh, Feng Hou, Xiaoyun Jia, and Ruili Wang. 2023. Phaseperturbation: Speech data augmentation via phase perturbation for automatic speech recognition. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia Workshops*, MMAsia '23 Workshops, Association for Computing Machinery, New York, NY, USA.

Li, Xinjian, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R. Mortensen, Graham Neubig, Alan W Black, and Florian Metze. 2020. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253.

Mennen, Ineke. 2015. *Beyond Segments: Towards a L2 Intonation Learning Theory*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Menéndez-Pidal, Xavier, James B. Polikoff, Shirley M. Peters, Jennie E. Leonzio, and H. T. Bunnell. 1996. The nemours database of dysarthric speech. In *4th International Conference on Spoken Language Processing (ICSLP 1996)*, pages 1962–1965.

Mortensen, David R., Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori S. 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, ACL.

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

Pang, Xudong, Aishan Wumaier, Wenwen Lu, Silajiaihemaiti Ruzemaimaiti, and Ligong Lei. 2026. A multi-task hierarchical deep reinforcement network approach for word level pronunciation assessment. *Expert Systems with Applications*, 317:131930.

Park, Daniel S., William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Interspeech 2019*, pages 2613–2617.

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

Rudzicz, Frank, Aravind Kumar Namasivayam, and Talya Wolff. 2012. The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Lang. Resour. Eval.*, 46(4):523–541.

Ryu, Hyuksu, Sunhee Kim, and Minhwa Chung. 2012. Comparing transcription agreement on non-native English speech corpus between native and non-native annotators. In *Interspeech 2012*, pages 2366–2369.

Sener, Ozan and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 525–536, Curran Associates Inc., Red Hook, NY, USA.

Shahin, Mostafa, Julien Epps, and Beena Ahmed. 2025. Phonological level wav2vec2-based mispronunciation detection and diagnosis method. *Speech Communication*, 173:103249.

Stevens, Kenneth N. 2000. *Acoustic Phonetics*. MIT Press.

Stouten, Frederik and Jean-Pierre Martens. 2006. Speech recognition with phonological features: some issues to attend. In *Interspeech 2006*, pages paper 1081–Mon2BuP.4.

Strömbergsson, Sofia, Katarina Holm, Jens Edlund, Tove Lagerberg, and Anita McAllister. 2020. Audience response system-based evaluation of intelligibility of children's connected speech – validity, reliability and listener differences. *Journal of Communication Disorders*, 87:106037.

Tadavarthy, Harsha Veena, Austin Jones, and Margaret E. L. Renwick. 2024. Phonological Feature Detection for US English using the Phonet Library. In *Interspeech 2024*, pages 1515–1519.

Thienpondt, Jenthe, Geoffroy Vanderreydt, Abdessalem Hammami, and Kris Demuynck. 2025. Weakly supervised phonological features for pathological speech analysis. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 1–5, IEEE.

Xu, Qiantong, Alexei Baevski, and Michael Auli. 2022. Simple and effective zero-shot cross-lingual phoneme recognition. In *Interspeech 2022*, ISCA.

Yang, Mu, Kevin Hirschi, Stephen D Looney, Okim Kang, and John HL Hansen. 2022. Improving mispronunciation detection with wav2vec2-based momentum pseudo-labeling for accentedness and intelligibility assessment. *arXiv preprint arXiv:2203.15937*.

Zhao, Guanlong, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018. L2-arctic: A non-native english speech corpus. In *Interspeech 2018*, page 2783–2787.

Appendix A: Additional Inadequate IPA-ARPAbet Articulatory Mappings

Table A.1
Language-Specific missing or ambiguous mappings between ARPAbet, IPA, and L2-ARCTIC annotations.

ARPA	IPA	L2-Arctic Substitute	Example	Comments
-	ħ	HH	ħamˈmaːm	Arabic pharyngeal
-	ʂ	SH	ʂɚ	Mandarin retroflex fricative
-	k^h	K HH	k^haːnaː	Hindi aspiration
-	ɟ	Y	kamɟi	Korean nasal
-	β	B	aβa	Spanish bilabial fricative
Q	ʔ	-	baʔ	Vietnamese glottal

Table A.2
Diphthongs and IPA symbols with diacritics

Index	ARPA	L2-ARCTIC IPA	Comments
5	AW	aʊ	diphthong
7	AY	aɪ	diphthong
9	CH	tʃ	affricate
13	ER	ɝ	rhotic allophone of ɜ
14	EY	eɪ	diphthong
20	JH	dʒ	affricate
22	L	ɫ	velarized allophone of l
26	OW	oʊ	diphthong
27	OY	ɔɪ	diphthong

Table A.3
Phonetic distinctions collapsed by the L2-ARCTIC ARPAbet inventory.

ARPA	IPA	L2-Arctic Substitute	Example	Comments
DX	ɾ	T	bottle /ˈbɑɾl/	approximation
EL	l̩	AH L	bottle /ˈbɑɾl̩/	closure
EM	m̩	AH M	rhythm /ˈrɪðm̩/	closure
EN	n̩	AH N	button /ˈbʊtn̩/	closure
NX	ɾ̃	N	winner /ˈwɪɾ̃ə/	nasalized
WH	ʍ	W	why /ʍaɪ/	approximation
AXR	ɚ	ER	forward /ˈfɔrwɚd/	rhotic approximation
IX	ɨ	IH	rabbit /ˈræbɨt/	approximation
UX	ʉ	UH	dude /dʉd/	approximation

Appendix B: Additional Evaluations and Results

We detail PERs for several other ablation combinations of architectures, training strategy, and augmentation presence below.

Table B.1
Change in errors (Best baseline to best model).

Metric	Best Baseline	Best HMTL	$\Delta\%$
PER (%)	14.46	13.47	-6.8%
Substitutions (within-class)	2889	2732	-5.4%
Deletions	705	520	-26.2%
Insertions	495	501	$+1.2\%$
Cross-class subs.	117	165	$+41.0\%$

Table B.2
Phoneme error types by model.

Model	Substitutions	Deletions	Insertions	Cross-class
Wav2Vec2Phoneme	3143	523	642	122
WavLM-base-plus	3003	776	499	121
WavLM + Aug	3055	673	555	123
WavLM + MPL	3067	657	504	123
WavLM + Aug. + MPL	3006	705	495	117
Parallel MTL	2978	381	745	207
Parallel MTL + Aug.	2914	414	715	190
Parallel MTL + Aug. + MPL	2871	365	766	207
Hierarchical MTL	2887	567	491	196
Hierarchical MTL + Aug.	2860	457	542	179
Hierarchical MTL + Aug. + MPL	2793	534	478	173
Hierarchical MTL + CA	2746	542	494	173
Hierarchical MTL + CA + Aug.	2699	576	488	169
Hierarchical MTL + CA + Aug. + MPL	2732	520	501	165

Table B.3
Phoneme feature errors by model.

Model	Cross-class	Place	Manner	Voicing	Height	Backness	Roundedness
Wav2Vec2Phoneme	122	716	603	810	1105	431	286
WavLM-base-plus	121	652	549	769	1094	441	275
WavLM + Aug	123	665	565	794	1087	444	290
WavLM + MPL	123	640	540	830	1074	436	301
WavLM + Aug + MPL	117	679	570	789	1044	429	275
Parallel MTL	207	715	627	786	1067	445	298
Parallel MTL+Aug.	190	690	609	771	1042	440	291
Parallel MTL+Aug+MPL	207	686	602	769	1021	429	283
Hierarchical MTL	196	669	576	798	1048	412	269
Hierarchical MTL+Aug.	179	672	578	734	1062	420	264
Hier MTL+Aug+MPL	173	666	555	717	1017	415	282
Hier MTL+CA	173	646	551	743	988	411	266
Hier MTL+CA+Aug.	169	648	538	709	976	411	252
Hier MTL+CA+Aug+MPL	165	657	548	709	985	404	262