Dialect ASR based on Multi-View Pseudo-Parallel Augmentation and Noise-Robust Contrastive Learning

Jianing Zhou
National Laboratory of Pattern Recognition, Institute of Automation, CAS
Beijing, 100083, China
jianing.zhou@ia.ac.cn

Ziheng Zeng
University of Illinois Urbana-Champaign
Urbana, IL 61801, USA
zzeng13@illinois.edu

Hongyu Gong
Facebook AI
Seattle, WA 98109, USA
hygong@fb.com

Suma Bhat
University of Illinois Urbana-Champaign
Urbana, IL 61801, USA
spbhat2@illinois.edu

Abstract

Recent advancements in Automatic Speech Recognition (ASR) have significantly improved performance for high-resource languages, yet ASR systems still struggle with low-resource languages and dialects due to the scarcity of annotated training data. To address this challenge, we propose a novel framework that enhances the utilization of limited data through a two-pronged approach: (1) a novel data augmentation technique based on a multi-view pseudo-parallel strategy, and (2) a robust multi-level contrastive learning framework capable of jointly leveraging semantic and dialect-specific attributes to improve model effectiveness under noisy conditions. Our approach effectively generalizes across dialectal variations when dialect data are non-parallel, allowing acquisition of shared linguistic structures and dialectal distinctions. Furthermore, our method could be simply transferred to scenarios where dialect data are parallel. We validate our method by adapting the Whisper-large model for Swiss German, Chinese and Arabic dialect ASR tasks, demonstrating substantial performance gains. Specifically, our framework achieves up to a 48.59% reduction in WER and 23.16% reduction in CER under non-parallel conditions, outperforming state-of-the-art baselines. These results highlight the robustness and adaptability of our approach for low-resource dialectal ASR. The data and codes are provided at `https://github.com/zhjjn/DialectASR.git`.

1 Introduction

Recent advances in neural architectures have drastically enhanced the capabilities of automatic speech recognition (ASR) systems, particularly for high-resource languages, leading to widespread use in various applications [1, 2, 3, 4]. However, the performance of these systems declines sharply when applied to low-resource languages and regional dialects, which are often underrepresented in training datasets [5, 6]. Dialects pose unique challenges due to their diverse pronunciation patterns shaped by factors such as regional influences, speaking styles, and sociolinguistic characteristics, including

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

the educational and cultural background of the speakers. The high cost and time requirements for annotating such speech data exacerbate the issue, creating a critical bottleneck for developing reliable ASR models tailored to dialectal speech.

To address these challenges, prior work has studied various approaches to improve ASR systems in dialectal and accented speech contexts. Methods such as multi-accent acoustic models with shared foundational layers and accent-specific components, as well as fine-tuning pre-trained models with limited accent-specific data, have shown potential [6, 7, 8, 9, 10]. Other studies have explored the use of dialect embeddings to improve recognition accuracy [11, 12]. Despite these advancements, most of these methods require specific features (abstracted as embeddings) and datasets restricting their general applicability in low-resource settings.

These limitations motivate us to leverage inter-dialect relationships for more efficient learning. Specifically, we investigate the following research questions: (1) Can features from similar or semantically parallel instances across dialects aid learning? (2) Is it possible to extract shared features from such instances? (3) Do these shared features improve ASR for unseen dialects? Guided by these questions, we address these limitations with a framework that utilizes available data resources more efficiently so as to make it suitable for truly low-resource scenarios. The key innovation lies in a novel data augmentation technique based on a multi-view strategy designed for low-resource contexts and a novel contrastive learning method based on multi-level contrastive learning to harness the commonalities among and the distinctions between different dialects. Unlike traditional augmentation methods that operate at the dataset level by adding more instances to the datasets (e.g., [13, 14]) or transforming instances to different formats (e.g., [15]), our proposed data augmentation strategy works at the instance level without relying on acquiring additional instances. For each training instance, we provide two other samples with different levels of parallelism under different scenarios, together creating complementary perspectives of the same data point. The augmented training instances are then used to effectively harness the commonalities among and the distinctions between different dialects.

More specifically, under a non-prallel scenario where semantically parallel instances from different dialects are not available in most practical scenarios, which is different from previous works [16], we provide each training instance with two additional samples that share different aspects of parallelism to enable multiple views of the data point that together offer related nuanced information. These paired instances are then processed in a noise-robust multi-level contrastive learning setting due to the existence of the pseudo-parallel instance. Our method enables ASR systems to generalize to previously unseen dialectal variations. Under the non-parallel scenario, the model gains a nuanced understanding of shared and distinctive attributes of dialects by analyzing the interrelations and variations between the original and augmented “views" of the data.

To evaluate our approach, we utilize Whisper-large [4], known for its strong performance on standard benchmarks such as LibriSpeech [17], as an ideal testbed. Using large dialectal datasets spanning different language families from different regions—the SwissDial dataset, the KeSpeech dataset [18] and the Casablanca dataset [19]—we demonstrate that our approach achieves notable improvements over finetuning in the non-parallel scenario. For example, In the non-parallel case, we observe an absolute reduction in WER by 48.59% and CER by 23.16% compared to the unmodified Whisper model on Swiss German dialects and a reduction in CER by 11.71% on Chinese dialects. To evaluate our method under parallel scenario, we also demonstrate our method’s performance on the SwissDial dataset which provides the semantically parallel instances from different dialects. Furthermore, our method consistently outperforms other non-trivial baselines on unseen dialects in both scenarios, highlighting its robustness and generalizability to diverse linguistic contexts.

Our contributions could be summarized as follows:

- We propose a novel framework that enhances the utilization of limited data for ASR for low-resource languages and dialects.
- We propose a novel data augmentation technique based on a multi-view pseudo-parallel strategy, which diversifies the limited training instances by incorporating information from related dialects.
- We propose a robust multi-level contrastive learning framework capable of jointly leveraging semantic and dialect-specific attributes to improve model effectiveness under noisy conditions.

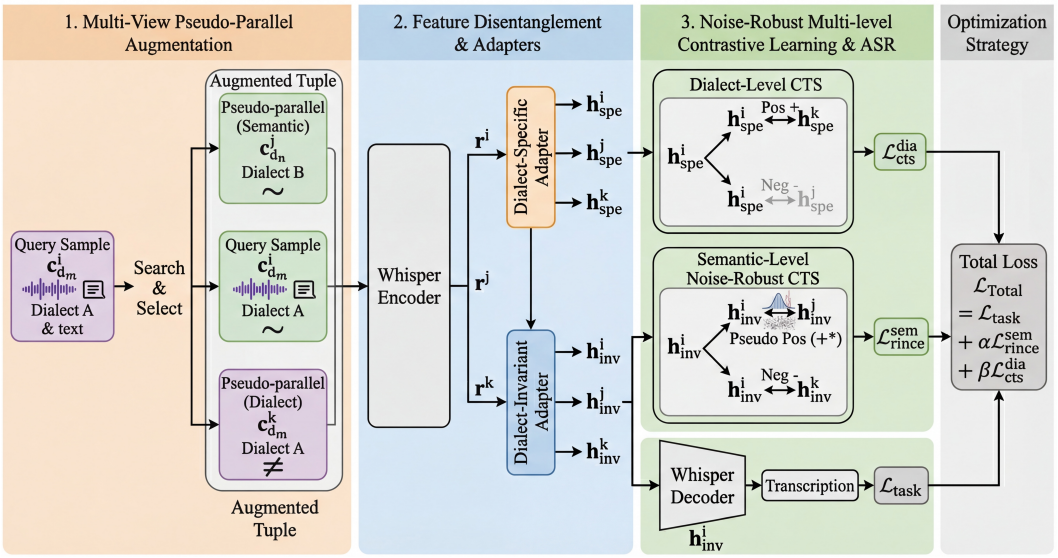

Figure 1: For each training instance $c^i_{d_m}$, we construct a multi-view tuple $\left(c^i_{d_m}, c^j_{d_n}, c^k_{d_m}\right)$ containing the query utterance, a semantically similar pseudo-parallel utterance from another dialect (selected via BERT-based transcript similarity), and a semantically dissimilar utterance from the same dialect. The tuple is encoded by a Whisper encoder to obtain representations $(r^i_{d_m}, r^j_{d_n}, r^k_{d_m})$, which are then mapped by a dialect-specific adapter E_{spe} and a dialect-invariant adapter E_{inv} into h_{spe} and h_{inv}, respectively. Training combines (i) dialect-level contrastive learning on h_{spe} to enforce dialect separation, (ii) semantic-level noise-robust contrastive learning (RINCE) on h_{inv} treating pseudo-parallel pairs as noisy positives, and (iii) the standard Whisper decoding ASR loss.

- Our proposed method achieves notable improvements across different languages and different datasets under a non-parallel scenario which could also be transferred to a parallel scenario.

2 Related Work

Pre-trained ASR models and low-resource adaptation. Large-scale pre-training has substantially improved ASR performance for high-resource languages, and modern architectures and training paradigms have enabled robust end-to-end recognition in diverse application settings [1, 2, 3, 4]. However, transferring these gains to low-resource languages and regional dialects remains challenging due to limited annotated data and systematic acoustic–phonetic shifts that are not well covered in pre-training corpora [5, 6]. Consequently, a common practice is to adapt a strong pretrained backbone (e.g., Whisper) via fine-tuning on the target dialect, but the effectiveness of direct fine-tuning is often constrained by data scarcity and dialect variability.

Dialect and accent modeling. A line of work improves recognition on accented and dialectal speech by explicitly modeling dialect variation. Representative approaches include multi-accent acoustic models that share a common backbone while allocating accent-specific components, as well as fine-tuning strategies that incorporate limited accent-specific supervision [6, 7, 8, 9, 10]. Another family of methods introduces dialect embeddings (or related side information) to condition the acoustic model and improve recognition accuracy [11, 12]. While effective when dialect identity is known and sufficient dialect coverage exists, these approaches often rely on additional features (abstracted as embeddings) and/or curated datasets, which can limit their applicability in truly low-resource scenarios where coverage is sparse and dialectal data are heterogeneous.

Positioning of our approach. Overall, our work complements existing dialect and accent adaptation methods by targeting a practically relevant setting where dialect data are *non-parallel*. We introduce an instance-level multi-view construction that exposes the model to two complementary comparisons, namely "similar semantics across dialects" and "different semantics within the same dialect", and we

optimize a multi-level contrastive objective that encourages functional specialization into a semantic space with reduced dialect predictability and a dialect-specific space. By combining pseudo-parallel augmentation with noise-robust contrastive learning on top of a strong pretrained ASR backbone [4], our approach aims to improve generalization under limited dialect supervision without relying on curated parallel corpora or additional hand-designed dialect features.

3 Proposed Method without Parallel Data

As mentioned previously, data resources with semantically parallel examples from different dialects are not available in most of the practical scenarios. Within a Whisper-based ASR system, our primary focus is on the data utilized for training. Considering the issue of limited data for low-resource dialects, it is crucial to maximize the use of available data. Therefore, we propose a multi-view data augmentation method relying on pseudo-parallel instances which to provide richer information. Furthermore, to better learn similarities and discrepancies in terms of semantic and dialect features, we propose a noise-robust multi-level contrastive learning scheme, described below in detail.

3.1 Multi-view Data Augmentation

Similarly to acquiring cross-modality features for ASR, cross-dialect features are traditionally acquired by aligning semantically parallel instances from different dialects [20]. Under low-resource scenarios, this requirement can be very difficult to obtain; it may be realistic to obtain audio samples and their transcriptions for each dialect, not necessarily with a one-to-one correspondence in terms of semantic equivalence. Thus, for any given instance within a specific dialect, it may possible to identify pseudo-parallel instances in other dialects that exhibit certain degrees of lexical or partial semantic overlap. These instances, despite their variances, may offer a more enriched source of cross-dialectal information compared to not using them. This prompts us to study a multi-view pseudo-parallel data augmentation approach that synthesizes additional training pairs even when the dataset is not parallel across dialects.

Let $\mathcal{D} = \{d_1, \ldots, d_M\}$ denote a set of M dialects. The training data are organized as

$$\mathcal{C} = \{\mathcal{C}_{d_1}, \ldots, \mathcal{C}_{d_M}\}, \qquad \mathcal{C}_{d_m} = \{c^i_{d_m}\}_{i=1}^{N_m}, \tag{1}$$

where each instance $c^i_{d_m}$ consists of an utterance $x^i_{d_m}$ and its transcription $y^i_{d_m}$. We use d_m (resp. d_n) to index dialects and i, j, k to index instances. Traditional cross-dialect feature learning often relies on aligning semantically parallel utterances across dialects, which is difficult under low-resource conditions. Instead, for each anchor instance $c^i_{d_m}$, we augment it into a *multi-view* tuple

$$\big(c^i_{d_m},\ c^j_{d_n},\ c^k_{d_m}\big), \tag{2}$$

where $d_n \neq d_m$. This tuple provides two complementary "views" of the same anchor: (i) a cross-dialect sample with *(approximately) similar semantics* and (ii) an in-dialect sample with *different semantics*. The resulting tuple is then used to jointly learn semantic commonalities and dialectal distinctions. More specifically, for a given training instance $c^i_{d_m}$ from dialect d_m, we find one semantically pseudo-parallel instance $c^j_{d_n}$ from another dialect d_n, and one semantically different instance $c^k_{d_m}$ from the same dialect d_m, respectively.

Since perfectly parallel instances may not exist, we identify a pseudo-parallel utterance $c^j_{d_n}$ by selecting the most semantically similar example in a different dialect based on contextual representations of transcriptions obtained from a mBERT model. Let $g(\cdot)$ be the mBERT embedding of a transcription and $\mathrm{sim}(\cdot,\cdot)$ be a similarity measure (e.g., cosine similarity). We select

$$(d_n, j) \in \arg \max_{\substack{d \in \mathcal{D}\setminus\{d_m\} \\ c^\ell_d \in \mathcal{C}_d}} \mathrm{sim}\big(g(y^i_{d_m}),\ g(y^\ell_d)\big). \tag{3}$$

This choice encourages lexical or partial semantic overlap across dialects while avoiding the need for explicit alignment.

To form the same-dialect contrasting view, we sample a semantically different utterance $c^k_{d_m}$ from the same dialect:

$$k \in \arg \min_{\substack{l \\ c^\ell_{d_m} \in \mathcal{C}_{d_m}}} \mathrm{sim}\big(g(y^i_{d_m}),\ g(y^\ell_{d_m})\big). \tag{4}$$

Table 1: The statistics of the audio data.

Datasets	SwissDial	KeSpeech	Casablanca
Non-Parallel Scenario Training Split			
Total Audio(h)	1.77	13.95	32.76
Average Utterance(s)	3.51	3.68	5.20
Average Number of Words per Utterance	10.40	11.21	14.38

This approach of augmenting each training instance $c^i_{d_m}$ into a tuple $(c^i_{d_m}, c^j_{d_n}, c^k_{d_m})$ now exposes the model to two distinct views of a given instance: one with (somewhat) similar semantics but in a different dialect, and the other with differing semantics but in the same dialect. As a result, we hypothesize that the model is better positioned to capture semantic nuances and generalize to unseen dialects by leveraging the learned features from known dialects.

Transcript embeddings use mBERT `[CLS]` features; embeddings and tuples are computed once, offline, before ASR training.

3.2 Noise-robust Multi-level Contrastive Learning

Given the input tuple $(c^i_{d_m}, c^j_{d_n}, c^k_{d_m})$ obtained by our proposed data augmentation method, we apply (i) a Whisper encoder to extract acoustic representations, (ii) two adapters to learn functionally specialized dialect-specific and dialect-invariant projections, and (iii) multi-level contrastive losses to supervise both spaces.

First, we encode each utterance with the Whisper encoder E_{wh}:

$$r^\ell_d = E_{\text{wh}}(x^\ell_d), \qquad (\ell, d) \in \{(i, d_m), (j, d_n), (k, d_m)\}, \tag{5}$$

where r^ℓ_d is a frame-level sequence of hidden representations.

Then we introduce two lightweight feed-forward adapters: a dialect-specific adapter E_{spe} and a dialect-invariant adapter E_{inv}. The adapters encourage functional specialization: the invariant branch preserves linguistic content while reducing dialect predictability, whereas the specific branch retains dialect-discriminative information. Neither objective guarantees complete factor separation. Thus, for a given instance in a dialect, we first use an adapter to extract their dialect-specific representation:

$$\mathbf{h}^i_{d_m,\text{spe}} = E_{\text{spe}}(\mathbf{r}^i_{d_m}), \mathbf{h}^j_{d_n,\text{spe}} = E_{\text{spe}}(\mathbf{r}^j_{d_n}), \mathbf{h}^k_{d_m,\text{spe}} = E_{\text{spe}}(\mathbf{r}^k_{d_m}) \tag{6}$$

where E_{spe} is the dialect-specific adapter and $\mathbf{h}^i_{d_m,\text{spe}}, \mathbf{h}^j_{d_n,\text{spe}}, \mathbf{h}^k_{d_m,\text{spe}}$ is intended to capture features unique to dialect.

To ensure that the dialect-specific representations for different dialects contain distinct dialect-related information, we treat the extracted dialect-specific representation for different dialects as negative pairs. We then apply contrastive learning at the dialect level to learn the dialect-specific representations.

$$\mathcal{L}^{\text{dia}}_{\text{cts}} = \log \frac{f(\mathbf{h}^i_{d_m,\text{spe}}, \mathbf{h}^k_{d_m,\text{spe}})}{f(\mathbf{h}^i_{d_m,\text{spe}}, \mathbf{h}^j_{d_n,\text{spe}}) + f(\mathbf{h}^k_{d_m,\text{spe}}, \mathbf{h}^j_{d_n,\text{spe}})}$$

where mean-pooling is applied over the frame-level features of adapter outputs to obtain fixed-dimensional utterance-level representations and $f(\cdot, \cdot) > 0$ is a cosine similarity function over utterance-level representations.

In contrast, the dialect-invariant branch encourages semantic consistency across dialects without requiring the shared encoder itself to discard dialect information. Therefore, for each tuple, we employ another adapter to learn the hidden dialect-invariant representations from the Whisper representations:

$$\mathbf{h}^i_{d_m,\text{inv}} = E_{\text{inv}}(\mathbf{r}^i_{d_m}), \mathbf{h}^j_{d_n,\text{inv}} = E_{\text{inv}}(\mathbf{r}^j_{d_n}), \mathbf{h}^k_{d_m,\text{inv}} = E_{\text{inv}}(\mathbf{r}^k_{d_m}).$$

where E_{inv} is the dialect-invariant adpater and $\mathbf{h}^i_{d_m,\text{inv}}, \mathbf{h}^j_{d_n,\text{inv}}, \mathbf{h}^k_{d_m,\text{inv}}$ aims to preserve semantic information common across dialects and suppress dialect idiosyncrasies.

To ensure consistency between dialects, we treat dialect-invariant representations derived from matching pseudo-parallel examples from different dialects as positive pairs, as they contain related

semantic information. Conversely, dialect-invariant representations from different examples within the same batch are considered negative pairs. We then apply contrastive learning at the semantic level to learn these dialect-invariant representations.

$$\mathcal{L}_{\text{cts}}^{\text{sem}} = -\log \frac{f(\mathbf{h}_{d_m,\text{inv}}^{i}, \mathbf{h}_{d_n,\text{inv}}^{j})}{f(\mathbf{h}_{d_m,\text{inv}}^{i}, \mathbf{h}_{d_m,\text{inv}}^{k}) + f(\mathbf{h}_{d_n,\text{inv}}^{j}, \mathbf{h}_{d_m,\text{inv}}^{k})} \tag{7}$$

Table 2: Overall performance comparison of different models on the Swiss German dialects under non-parallel scenario.

	sg		vs		lu		gr		be		zh		ag		bs		Average	
	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer
XLS-R 1B	64.01	32.42	70.61	33.83	73.15	34.76	63.62	32.33	74.83	35.54	64.72	32.21	79.53	36.41	64.57	32.93	69.38	33.8
Whisper	79.09	26.64	81.21	31.57	87.81	31.86	69.61	26.03	83.41	34.32	75.63	27.58	88.41	32.32	69.62	24.69	79.13	30.00
Dolphin-base	28.89	7.33	35.49	8.74	38.03	9.67	28.50	7.24	39.71	10.45	29.60	7.12	44.41	11.32	29.45	7.84	34.26	8.71
SS-kNN	31.11	8.73	37.71	10.14	40.25	11.07	30.72	8.64	41.93	11.85	31.82	8.52	46.63	12.72	31.67	9.24	36.48	10.11
Whisper-ft	49.46	34.93	60.79	39.97	60.45	38.64	48.87	34.67	63.69	40.54	51.58	34.65	66.45	40.39	50.22	35.41	54.47	37.41
+ SpecAugment	37.13	14.73	45.63	16.86	45.38	16.30	36.69	14.62	47.81	17.10	38.72	14.61	49.88	17.03	37.70	14.93	42.37	15.77
Ours (on XLS-R 1B)	23.98	5.10	29.83	6.35	32.08	7.18	23.64	5.02	33.57	7.87	24.61	4.92	37.73	8.64	24.48	5.56	28.74	6.33
Ours	**27.08**	**5.76**	**33.68**	**7.17**	**36.22**	**8.1**	**26.69**	**5.67**	**37.9**	**8.88**	**27.79**	**5.55**	**42.6**	**9.75**	**27.64**	**6.27**	**30.54**	**6.84**
- w/o sem	38.62	11.54	43.19	10.43	42.86	11.17	42.6	14.82	44.52	10.5	35.65	8.26	48.11	11.9	37.98	10.26	39.54	10.75
- w/o dia	37.65	9.02	41.96	8.93	42.79	9.39	39.37	9.04	42.61	9.63	36.03	7.03	48.9	11.5	35.1	7.92	38.84	9.56

Table 3: Overall performance comparison of different models on the KeSpeech dataset. Performance is evaluated based on CER

	Northeastern	Ji-Lu	Beijing	Jiang-Huai	Jiao-Liao	Lan-Yin	Mandarin	Zhongyuan	Southwestern	Ave.
XLS-R 1B	35.12	41.24	38.69	50.89	41.31	42.95	33.44	39.05	40.15	40.32
Whisper	21.68	39.67	35.31	52.55	42.06	48.55	21.65	40.35	42.64	37.53
Dolphin-base	11.54	17.66	15.11	27.31	17.73	19.37	9.86	15.47	16.57	16.74
SS-kNN	16.06	22.18	19.63	31.83	22.25	23.89	14.38	19.99	21.09	19.18
Whisper-ft	14.04	25.1	21.66	37.38	26.55	28.85	12.37	22.9	26.27	23.77
+ SpecAugment	12.24	21.88	18.88	32.59	23.15	25.15	10.78	19.97	22.91	20.84
Ours (XLS-R 1B)	12.25	19.20	16.31	30.15	19.28	21.14	10.35	16.71	17.96	18.15
Ours	**10.8**	**16.92**	**14.37**	**26.57**	**16.99**	**18.63**	**9.12**	**14.73**	**15.83**	**15.82**
- w/o semantic cts	14.09	25.64	21.21	37.57	25.81	27.86	12.61	22.03	26.41	22.32
- w/o dialect cts	11.36	17.72	16.21	25.04	19.43	23.01	8.55	18.6	15.55	17.08

Table 4: Overall performance comparison of different models on the Arabic dialects under non-parallel scenario.

	Morocco		Palestine		Jordan		UAE		Yemen		Algeria		Mauritania		Egypt		Average	
	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer
XLS-R 1B	82.65	58.87	63.33	50.79	61.25	48.43	68.03	51.55	85.10	60.19	85.66	60.44	91.39	67.31	64.24	52.72	75.21	41.33
Whisper	92.2	55.61	59.64	24.08	51.26	19.92	64.75	29.28	75.91	36.28	87.13	49.06	94.02	63.66	62.52	28.01	74.29	39.17
Dolphin-base	64.58	28.05	46.26	19.97	44.18	17.61	49.96	20.73	68.03	29.37	68.59	29.62	73.32	36.49	47.17	21.90	55.14	25.47
SS-kNN	65.73	30.89	47.41	22.81	45.33	20.45	52.11	23.57	68.18	32.21	68.74	32.46	74.47	39.33	48.32	24.74	56.29	28.31
Whisper-ft	72.63	33.41	53.65	21.14	53.81	20.71	61.0	23.63	73.25	30.91	76.96	33.79	84.92	45.81	58.09	26.78	62.83	28.13
+ SpecAugment	64.65	31.04	47.76	19.64	47.90	19.24	54.30	21.95	65.21	28.72	68.51	31.39	75.59	42.23	51.71	24.88	59.46	27.39
Ours (XLS-R 1B)	**61.25**	27.74	**43.01**	18.43	**41.05**	15.71	**47.45**	19.31	**63.57**	29.26	**64.10**	29.55	**69.50**	37.47	**43.87**	20.66	54.23	24.77
Ours	64.88	**24.08**	45.56	**16.0**	43.48	**13.64**	50.26	**16.76**	67.33	**25.4**	67.89	**25.65**	73.62	**32.52**	46.47	**17.93**	**53.39**	**20.23**
- w/o sem	68.36	26.02	46.88	16.4	46.08	14.56	52.17	17.02	67.71	25.98	70.67	27.05	75.85	34.47	51.35	19.66	55.15	21.09
- w/o dia	68.74	26.05	46.59	16.37	45.18	14.4	51.69	17.13	67.07	25.45	68.43	25.98	76.08	34.65	50.62	19.51	54.89	21.04

However, our input tuples being pseudo-parallel instances extracted based on semantic similarity, can inject noise in the training tuples. To counter this and mitigate the negative influence caused by pseudo-positive pairs, we incorporate noisy contrastive learning [21]:

$$\mathcal{L}_{\text{rince}}^{\text{sem}} = -(1-\lambda)e^{s^+} + \lambda(e^{s_1^-} + e^{s_2^-}) \tag{8}$$

where $s^+ = f(\mathbf{h}_{d_m,\text{inv}}^{i}, \mathbf{h}_{d_n,\text{inv}}^{j})$, $s_1^- = f(\mathbf{h}_{d_m,\text{inv}}^{i}, \mathbf{h}_{d_m,\text{inv}}^{k})$ and $s_2^- = f(\mathbf{h}_{d_n,\text{inv}}^{j}, \mathbf{h}_{d_m,\text{inv}}^{k})$.

Similarly, mean-pooling is applied over the frame-level features of adapter outputs to obtain fixed-dimensional utterance-level representations for obtaining the similarities and calculating the contrastive loss and and $f(\cdot,\cdot) > 0$ is a cosine similarity function over utterance-level representations. Finally, the overall loss function is the sum of the above losses, which we minimize during model training:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \alpha\mathcal{L}_{\text{rince}}^{\text{sem}} + \beta\mathcal{L}_{\text{cts}}^{\text{dia}}$$

where $\mathcal{L}_{\text{task}}$ refers to the standard ASR cross-entropy loss.

The shared encoder r need not be dialect-invariant: the two losses supervise separate projections. Only frame-level h_{inv} conditions the decoder; h_{spe} is a training-only auxiliary branch. By contrasting utterances with different content but the same dialect, the auxiliary objective makes dialect variation explicit through gradients to the shared encoder, complementing cross-dialect semantic alignment. This provides an inductive bias for specialization rather than a guarantee of disentanglement. Decoding requires neither dialect labels nor retrieval.

It should be noted that our method could be simply applied to scenarios where semantically parallel instances from different dialects are available under the parallel scenario by replacing the pseudo-parallel example from another dialect in our method with a truly parallel example from another dialect. Due to the space limitation, details are provided in the Section B.

4 Experiments

We present the experimental details, including the datasets used, the experimental setups, corresponding results, discussions and analyses.

4.1 Datasets

We use the SwissDial corpus [22] for Swiss German dialects, the KeSpeech corpus [18] for Chinese dialects, and the Casablanca corpus [18] for Arabic dialects. Summarization of the corpora is shown in Table 1 and about their processing in the following:

SwissDial. The SwissDial corpus [22] constitutes a comprehensive parallel dataset encompassing eight major Swiss German dialects (AG, BE, BS, GR, LU, SG, VS, ZH). We first selected only sentences spoken by speakers from all eight dialects to ensure balance across dialects. Then we randomly sampled 30% of the data to form a test set. During sampling, we ensured different recordings of the same sentence spoken by different speakers were assigned to the *same* set. This guarantees a strict separation between the training and test sets, differing from [20], and leading to a more realistic evaluation. Details of the number of utterances and hours of recordings are shown in Table 9. To adapt this resource for non-parallel scenario, we constructed the training set by randomly selecting a single dialect's audio segment for each Standard German transcript, thereby eliminating inter-dialect alignment. Details of the training split used under non-parallel scenario are shown in Table 9. The training split contains only 1.77 hours in total (0.17–0.29 hours per dialect), directly evaluating the sub-one-hour-per-dialect regime.

KeSpeech. The KeSpeech corpus [18] comprises 1,542 hours of Mandarin speech data collected from 27,237 speakers across 34 Chinese regions, encompassing both standard Mandarin and eight regional variants. This corpus is only used under the non-parallel scenario. To avoid the class imbalance we randomly sample 2538 examples, the number of examples from the dialect with the fewest examples, for every dialect. As no predefined data partition was provided, we implemented a systematic split by reserving 30% of randomly selected samples for testing, ensuring proportional representation of dialectal varieties. Statistical breakdowns of the resultant dataset are comprehensively outlined in Table 11.

Casablanca. The Casablanca corpus [18] comprises 48 hours of Arabic speech data collected from 8 dialects, encompassing Algerian, Egyptian, Emirati, Jordanian, Mauritanian, Moroccan, Palestinian, and Yemeni. This corpus is only used under the non-parallel scenario. As official data partition was provided and only validation and test set are released, we utilize the validation set for training and test set for test. Statistical breakdowns of the resultant dataset are comprehensively outlined in Table 10.

4.2 Experimental Setups

Our investigation employs the Whisper-large-v2 architecture, featuring 24 transformer layers [23]. Adapters are lightweight feed-forward layers for extracting dialect-specific and dialect-invariant features. Optimization utilizes the Adam algorithm [24] with a constant batch size of 32 and initial learning rate of 1e-5, modulated by a linear warmup schedule (10% of total steps) transitioning to polynomial decay. Regularization parameters α and β in the loss function were fixed at 0.01 throughout training. Results are averaged on 5 different runs. Comprehensive exploration of the

Table 5: Performance on SwissDial with different scales of training set. **Bold** refers to the performance on the unseen dialects.

	sg		vs		lu		gr		be		zh		ag		bs		Average	
	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer
Less data																		
Whisper-ft	57.09	25.91	73.13	31.36	**77.45**	**31.5**	59.37	26.35	69.56	30.47	57.7	25.85	**81.47**	**31.95**	56.95	26.41	63.75	27.93
Ours	29.34	6.31	38.1	8.1	**60.18**	**13.39**	29.65	6.28	41.41	9.7	32.59	6.46	**67.72**	**5.55**	29.09	6.61	38.53	8.72
Mini data																		
Whisper-ft	56.71	25.92	70.47	31.15	**76.21**	**31.09**	60.56	26.22	**71.82**	**31.27**	57.42	25.55	**79.95**	**31.85**	**59.2**	**26.7**	63.9	27.94
Ours	29.23	6.12	34.57	7.45	**59.39**	**13.46**	27.66	5.93	**57.82**	**14.06**	29.81	6.04	**65.31**	**14.78**	**39.86**	**8.93**	40.64	9.26

Table 6: Performance comparison of different methods on the KeSpeech dataset with different scales of training set.

	Northeastern	Ji-Lu	Beijing	Jiang-Huai	Jiao-Liao	Lan-Yin	Mandarin	Zhongyuan	Southwestern	Ave.
Less data										
Whisper-ft	17.28	25.52	23.16	**39.17**	**27.01**	29.87	14.93	23.74	27.23	25.19
Ours	11.59	17.95	16.08	**26.28**	**16.43**	19.01	9.04	16.64	16.02	16.37
Mini data										
Whisper-ft	**17.74**	**26.75**	24.43	**40.16**	**28.19**	31.16	14.84	24.65	26.18	25.88
Ours	**11.38**	**17.82**	16.33	**24.9**	**19.45**	23.46	8.81	18.79	15.5	17.21

Table 7: Performance comparison for Arabic dialects with different scales of training set.

	Morocco		Palestine		Jordan		UAE		Yemen		Algeria		Mauritania		Egypt		Average	
	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer
Less data																		
Whisper-ft	85.45	40.86	51.67	19.96	53.34	20.15	58.05	22.29	71.0	29.43	77.4	34.72	90.96	51.45	55.84	25.25	68.82	31.2
Ours	84.29	40.19	50.15	18.92	49.34	18.22	56.33	21.07	70.83	29.21	74.9	32.44	89.17	49.33	55.56	24.89	63.21	28.36
Mini data																		
Whisper-ft	87.72	44.06	49.79	19.11	50.13	18.83	56.12	21.53	78.41	33.47	89.69	42.83	90.24	51.55	53.82	24.18	70.16	32.63
Ours	88.64	44.15	48.54	18.42	48.44	17.69	55.17	21.18	76.51	31.87	89.3	42.46	90.19	50.65	54.16	24.32	65.58	30.24

influence of the hyper-parameters is left for future study. All configurations underwent 20-epoch training cycles to ensure convergence.

For baseline models, we choose the models with comparable sizes which could also be generalized to different languages including XLS-R [25], Whisper-large-v2 [4], Dolphin [26] and SS-kNN [27]. It should be noted that Dolphin and SS-kNN also utilized different data augmentation methods during their training stage.

4.3 Results and Discussions

4.3.1 Main Results

Our first experiment involves a zero-shot evaluation on the test sets of SwissDial and KeSpeech. As shown in Tables 2, 3 and 4, although different models shows a strong performance on standard benchmarks [4], their performance on the Swiss German, the Chinese and the Arabic dialects is notably poor, with word error rates (WER) exceeding 75% on Swiss German, a CER exceeding 37% on Chinese and a CER exceeding 39% on Arabic. This significant degradation highlights the challenges different audio models face when processing dialectal speech, especially in cases with substantial accent changes compared to the training data.

Then, we fine-tuned the best performing base model, Whisper-large-v2, directly on the SwissDial, KeSpeech and Casablanca dataset respectively under non-parallel scenario. As shown in Table 2, 3 and 4, fine-tuning leads to substantial improvements in ASR for all dialects. This improvement demonstrates the importance of incorporating in-domain speech data for enhancing the model's performance on low-resource ASR tasks. Despite the performance gains compared to the zero shot setting, we hypothesize that direct fine-tuning on the original dialect dataset does not fully exploit the available data, resulting in suboptimal performance.

To further enhance the model's performance, we applied our method on the best performing model, Whisper-large-v2. The results of using the proposed pipeline are summarized in Table 2, 2 and 3. We note significant gains over the fine-tuned baseline, reducing the WER to 30.54% and the CER to 6.84% on Swiss German, the CER on Chinese to 15.82% and the CER on Arabic to 20.23%. More importantly, our proposed model consistently achieves superior performance across all dialects for Swiss German, Chinese and Arabic, demonstrating its robust generalization capability not only on the dialect level but also on the language level. Since both the baseline and our approach share

Table 8: Additional analyses (%). (a) SwissDial component comparisons. (b) Frozen-representation dialect-probe accuracy on SwissDial (SD, 8-way), KeSpeech (Ke, 9-way), and Casablanca (Ca, 8-way).

(a) Component comparisons

Variant	WER	CER
Full model	30.54	6.84
w/o semantic loss	39.54	10.75
w/o dialect loss	38.84	9.56
InfoNCE	40.67	11.12
Random in-dialect	32.41	7.93
Whisper-ft + SpecAug.	42.37	15.77

(b) Dialect probing

Representation	SD	Ke	Ca
Chance ($1/K$)	12.5	11.1	12.5
r	64.7	49.3	60.2
h_{spe}	77.1	67.4	75.3
h_{inv}	21.3	15.6	19.7

the same task-specific fine-tuning configuration, we attribute the improvements to the richer audio representations facilitated by the augmented data and learned through multi-level contrastive learning. These representations allow the model to better adapt to the unique speech characteristics of the target dialects, thereby addressing the challenges of dialectal ASR in low-resource settings. These findings provide insights into our first research question and show how features drawn from similar or parallel instances from different dialects are helpful for dialect ASR.

4.3.2 Ablation Study

Removing either contrastive objective degrades recognition across all three datasets (Tables 2–4). Table 8(a) consolidates SwissDial comparisons: removing semantic or dialect supervision increases WER by 9.00 or 8.30 points, respectively, while replacing RINCE with InfoNCE adds 10.13 points. Random same-dialect selection has a smaller effect. These conditional effects are not additive, since tuple construction and the contrastive objectives are coupled.

4.3.3 Dialect Probing

We freeze the trained ASR model, mean-pool r, h_{spe}, and h_{inv} while masking padding, and fit identical multinomial logistic-regression dialect probes on a stratified held-out split, evaluating on disjoint utterances. Table 8(b) shows that h_{spe} increases dialect predictability while h_{inv} reduces it across all three datasets. Above-chance accuracy indicates residual dialect information; the results support functional specialization, not statistical independence or identifiable disentanglement.

4.3.4 Out-of-Domain Performance as a Function of Dialects in the Training Data

To further analyze the generalizability of our proposed method, we empirically evaluate our model in an out-of-domain scenario. We exclude different numbers of dialects from the training set, but keep the test set unchanged. Thus, the excluded dialects will be unseen and out-of-domain for the model fine-tuned on the resulting smaller training set. First, we exclude two dialects from the training set including Luzern (LU) and Aargau (AG) which are least similar to the standard language for Swiss German under both scenarios, Jiao-Liao (JL) and Jiang-Huai (JH) for Chinese and Morocco and Mauritania for Arabic. This setup is referred to as **Less data**. Then, we exclude two more dialects from the training set including Basel (BS) and Bern (BE) for Swiss German, Northeastern (Nt) and Ji-Lu (JiLu) for Chinese and Algeria and Yemen for Arabic, resulting in a 50% reduction in training size which is **Mini data**.

As shown in Tables 5, 6, and 7 our proposed approach results in a superior performance compared to direct fine-tuning. The aggregate advantage persists across all three datasets under both reduced-coverage settings, although improvements are not uniform for every individual dialect. The performance on the unseen dialects occurs while still maintaining satisfactory performance on the seen dialects under both setups. This helps answer the third research question that the common features learned during training help with related but unseen dialects but the benefits are limited, which calls for further improvement on our framework.

Reported averages cover the unchanged test sets, including seen and excluded dialects; "unseen" means absent during task-specific fine-tuning.

To validate the effectiveness of our method under the parallel scenario, we summarize the results of using our proposed pipeline under the parallel scenario in Table 13.

4.3.5 mBERT Quality

On SwissDial's full parallel training set, mBERT `[CLS]` achieves 87.1% top-1 cross-dialect retrieval precision. In the non-parallel ASR setting, replacing 50% of retrieved cross-dialect views with random samples increases WER/CER from 30.54/6.84 to 43.19/18.93 (full results in Appendix C). Thus, structured matching remains important: RINCE mitigates pair noise but does not eliminate dependence on meaningful cross-dialect semantic overlap.

4.3.6 RINCE Influence

Replacing RINCE with InfoNCE while retaining the remaining framework increases SwissDial WER/CER from 30.54/6.84 to 40.67/11.12. The same trend holds on KeSpeech and Casablanca (Appendix C), supporting noise-robust optimization of mined pseudo-positive pairs.

4.3.7 Backbone Generalization

To explore the generalization ability of our proposed method, we applied the same framework to XLS-R model on all three datasets including SwissDial, KeSpeech and Casablanca. The results are reported in Tables 2, 3 and 4. Based on the experiments, our method could generalize to other backbones, which shows that our proposed method could effectively utilizes dialectal variations to enhance model generalization to dialects without requiring additional annotated data.

5 Conclusion

In this work, we explored the challenges of ASR for low-resource dialects by introducing a novel framework including a novel multi-view data augmentation method and a multi-level contrastive learning framework. Our approach effectively utilizes dialectal variations to enhance model generalization to unseen dialects without requiring additional annotated data for those dialects. By integrating semantic and dialect-specific contrastive learning, our model significantly improves recognition accuracy, reducing WER and CER across Swiss German, Chinese and Arabic dialects under non-parallel scenario and even out-of-domain scenarios. Extensive experiments demonstrate that (1) Related features provided by similar instances from different dialects are helpful, (2) multi-level contrastive learning more effectively captures shared features than simpler baselines, as confirmed by ablation studies, and (3) these common features learned during training are helpful for recognizing unseen dialects given that our methods achieve improvements even in out-of-domain scenarios but the benefits are limited, which calls for future study.

6 Acknowledgments

This work was supported in part by the National Science Foundation and the Institute of Education Sciences under Grant #2229612. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the U.S. Department of Education. This work was also supported in part by Soliciting Projects for the Shaanxi Science and Technology Development Plan under Grant No.2026CY-ZDGG-102.

References

[1] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[3] F. Weninger, M. Gaudesi, M. A. Haidar, N. Ferri, J. Andrés-Ferrer, and P. Zhan, "Conformer with dual-mode chunked attention for joint online and offline asr," *arXiv preprint arXiv:2206.11157*, 2022.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[5] Y. Zhang, Y. Zhang, B. M. Halpern, T. Patel, and O. Scharenborg, "Mitigating bias against non-native accents," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2022, 2022, pp. 3168–3172.

[6] L. Lonergan, M. Qian, N. N. Chiaráin, C. Gobl, and A. N. Chasaide, "Towards dialect-inclusive recognition in a low-resource language: are balanced corpora the answer?" *arXiv preprint arXiv:2307.07295*, 2023.

[7] N. Das, S. Bodapati, M. Sunkara, S. Srinivasan, and D. H. Chau, "Best of both worlds: Robust accented speech recognition with adversarial transfer learning," *arXiv preprint arXiv:2103.05834*, 2021.

[8] L. M. Arslan and J. H. Hansen, "A study of temporal features and frequency characteristics in american english foreign accent," *The Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 28–40, 1997.

[9] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 1. IEEE, 2003, pp. I–I.

[10] J. Eisenstein, V. Prabhakaran, C. Rivera, D. Demszky, and D. Sharma, "Md3: The multi-dialect dataset of dialogues," *arXiv preprint arXiv:2305.11355*, 2023.

[11] S. Li, B. Ouyang, D. Liao, S. Xia, L. Li, and Q. Hong, "End-to-end multi-accent speech recognition with unsupervised accent modelling," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6418–6422.

[12] Y. Qian, X. Gong, and H. Huang, "Layer-wise fast adaptation for end-to-end multi-accent speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2842–2853, 2022.

[13] A. Ullah, A. Ragano, and A. Hines, "Reduce, reuse, recycle: Is perturbed data better than other language augmentation for low resource self-supervised speech models," *arXiv preprint arXiv:2309.12763*, 2023.

[14] D. Mujtaba, N. R. Mahapatra, M. Arney, J. S. Yaruss, C. Herring, and J. Bin, "Inclusive asr for disfluent speech: Cascaded large-scale self-supervised learning with targeted fine-tuning and data augmentation," *arXiv preprint arXiv:2406.10177*, 2024.

[15] J. Zhou, Z. Zeng, H. Gong, and S. Bhat, "Enhancing language models with idiomatic reasoning," in *First Conference on Language Modeling*.

[16] ——, "Clasp: Cross-modal alignment using pre-trained unimodal models," in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 11 518–11 531.

[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[18] Z. Tang, D. Wang, Y. Xu, J. Sun, X. Lei, S. Zhao, C. Wen, X. Tan, C. Xie, S. Zhou *et al.*, "Kespeech: An open source speech dataset of mandarin and its eight subdialects," in *Thirty-fifth*

Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.

[19] B. Talafha, K. Kadaoui, S. Magdy, M. Habiboullah, C. Chafei, A. El-Shangiti, H. Zayed, M. Tourad, R. Alhamouri, R. Assi *et al.*, "Casablanca: Data and models for multidialectal arabic speech recognition," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 21 745–21 758.

[20] C. Sicard, V. Gillioz, and K. Pyszkowski, "Spaiche: Extending state-of-the-art asr models to swiss german dialects," in *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, 2023, pp. 76–83.

[21] C.-Y. Chuang, R. D. Hjelm, X. Wang, V. Vineet, N. Joshi, A. Torralba, S. Jegelka, and Y. Song, "Robust contrastive learning against noisy views," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 670–16 681.

[22] P. Dogan-Schönberger, J. Mäder, and T. Hofmann, "Swissdial: Parallel multidialectal corpus of spoken swiss german," *arXiv preprint arXiv:2103.11401*, 2021.

[23] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[24] D. P. Kingma and J. L. B. Adam, "A method for stochastic optimization. int. conf. learn," *Representations (ICLR)*, vol. 6, 2015.

[25] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," in *Proc. Interspeech 2022*, 2022, pp. 2278–2282.

[26] Y. Meng, J. Li, G. Lin, Y. Pu, G. Wang, H. Du, Z. Shao, Y. Huang, K. Li, and W.-Q. Zhang, "Dolphin: A large-scale automatic speech recognition model for eastern languages," *arXiv preprint arXiv:2503.20212*, 2025.

[27] S. Li, D. Wei, H. Shang, J. Guo, Z. Li, Z. Wu, Z. Rao, Y. Luo, X. He, and H. Yang, "Speaker-smoothed knn speaker adaptation for end-to-end asr," in *Proc. Interspeech 2024*, 2024, pp. 2390–2394.

A Dataset Statistics

Details of the datasets used are shown in Tables 9, 10, and 11.

Table 9: The statistics of the audio data in SwissDial dataset.

Swiss German	AG	BE	BS	LU	GR	SG	VS	ZH	Total
Non-Parallel Scenario Training Split									
Total Audio(h)	0.18	0.25	0.20	0.17	0.21	0.29	0.24	0.23	1.77
Average Utterance(s)	3.05	4.05	3.67	2.81	3.33	4.43	3.76	3.71	3.51
Average Number of Words per Utterance	9.94	10.21	10.98	10.55	10.71	10.39	10.03	10.57	10.40

Table 10: The statistics of the audio data in Casablanca dataset.

Arabic	Morocco	Palestine	Jordan	UAE	Yemen	Algeria	Mauritania	Egypt	Total
Non-Parallel Scenario									
Total Audio(h)	1.72	1.79	1.76	1.71	1.81	1.68	1.70	1.78	13.95
Average Utterance(s)	2.97	4.84	3.74	3.78	4.07	3.59	3.21	3.79	3.68
Average Number of Words per Utterance	11.38	13.34	10.54	11.13	11.51	9.75	10.63	11.94	11.21

Table 11: The statistics of the audio data in KeSpeech dataset.

	Northeastern	Ji-Lu	Beijing	Jiang-Huai	Jiao-Liao	Lan-Yin	Mandarin	Zhongyuan	Southwestern	Total
Non-Parallel Scenario										
Audio (h)	3.74	3.71	3.71	3.50	3.60	3.59	3.60	3.71	3.58	32.76
Utterance(s)	5.30	5.26	5.27	4.96	5.11	5.10	5.10	5.30	5.08	5.20
Number of Words	14.00	14.36	14.47	14.50	14.41	14.32	14.39	14.42	14.50	14.38

B Our Method With Parallel Data

As mentioned in the main context, our proposed method could also be easily transferred to the parallel scenario. Due to the fact that there are semantically parallel instances from different dialects under the parallel scenario, the pseudo-parallel example from another dialect in our method could be directly replaced with a truly parallel example from another dialect. Furthermore, the noise-robust contrastive learning could be replaced with a standard contrastive learning. Here we first introduce our multi-view data augmentation method that utilizes parallel instances from different dialects for learning commonalities and discrepancies between different dialects. Then, we introduce our multi-level contrastive learning to better disentangle semantic and dialect features and thus further enhance the effect of the multi-view data augmentation method.

B.1 Multi-view Data Augmentation

Formally, we consider a parallel dataset $C = \{(c_i^{d_1}, c_i^{d_2}, \ldots, c_i^{d_m}, \ldots, c_i^{d_M})\}_{i=1}^{N}$ where d_m represents dialect m, and $(c_i^{d_1}, c_i^{d_2}, \ldots, c_i^{d_m}, \ldots, c_i^{d_M})$ comprises semantically equivalent training instances expressed and enunciated in different dialects. For a given training instance $c_i^{d_m}$ from dialect d_m, we augment the data by extracting two semantically parallel instances, $c_i^{d_j}$ and $c_i^{d_k}$, from dialects d_j and d_k, respectively.

The selection of $c_i^{d_j}$ and $c_i^{d_k}$ is guided by specific criteria. First, $c_i^{d_j}$ is chosen based on the similarity of the dialect d_j to the standard language (in our case, German), which is measured by the BLEU scores between transcriptions in the standard language and transcriptions in dialects[1]. This ensures that d_j serves as a consistent linguistic "bridge" between all dialects, facilitating model generalization to other dialects. Second, $c_i^{d_k}$ is selected based on its similarity to the original instance $c_i^{d_m}$. Specifically, we identify d_k as the dialect that produces an instance most similar to $c_i^{d_m}$, as measured by the cosine similarity of utterance representations obtained from the Whisper model.

Such an augmentation of each training instance $c_i^{d_m}$ into a triplet $(c_i^{d_m}, c_i^{d_j}, c_i^{d_k})$, exposes the model to two distinct perspectives of the same semantic content, derived from different dialects. This multi-view augmentation enables the model to learn both the commonalities and the distinctions

[1]We leave other methods of measuring semantic similarity for future work.

in semantics among dialects, permitting it to adapt to unseen dialects by leveraging the learned similarities and features from the seen dialects.

B.2 Multi-level Contrastive Learning

Given the input tuple $(c_i^{d_m}, c_i^{d_j}, c_i^{d_k})$ obtained by our proposed data augmentation method, we project each into functionally specialized *dialect-invariant* and *dialect-specific* representations, with the same interpretation and limitations as in the non-parallel setting. The dialect-invariant representation captures the shared semantics across dialects, ideally remaining identical across parallel instances. To learn the dialect-invariant representation, we first apply a linear layer to the Whisper-generated representations:

$$\mathbf{h}^i_{d_m,\text{inv}} = E_{\text{inv}}(\mathbf{r}_i^{d_m}), \mathbf{h}^i_{d_j,\text{inv}} = E_{\text{inv}}(\mathbf{r}_i^{d_j}), \mathbf{h}^i_{d_k,\text{inv}} = E_{\text{inv}}(\mathbf{r}_i^{d_k}).$$

where $(\mathbf{r}_i^{d_m}, \mathbf{r}_i^{d_j}, \mathbf{r}_i^{d_k})$ are the representations obtained from Whisper. Besides, 'inv' means dialect-invariant, and E means adapter.

To ensure semantic consistency across dialects, we treat dialect-invariant representations derived from the parallel triplet examples from different dialects as *positive* pairs (owing to their shared semantics), while representations from different triplets within the same batch are considered *negative* pairs. We then apply contrastive learning at the semantic level to learn these dialect-invariant representations.

$$\mathcal{L}^{\text{sem}}_{\text{cts}} = -\log \frac{f(\mathbf{h}^i_{d_m,\text{inv}}, \mathbf{h}^i_{d_j,\text{inv}}) + f(\mathbf{h}^i_{d_m,\text{inv}}, \mathbf{h}^i_{d_k,\text{inv}})}{\sum_{d\in\{d_m,d_j,d_k\}} \sum_{j\neq i} f(\mathbf{h}^i_{d,\text{inv}}, \mathbf{h}^j_{d,\text{inv}})} \tag{9}$$

where f is the cosine similarity scoring function.

In contrast, dialect-specific representations are trained to emphasize dialect-discriminative information; this objective does not guarantee the complete removal of semantic content. To extract these representations, we employ another linear layer:

$$\mathbf{h}^i_{d_m,\text{spe}} = E_{\text{spe}}(\mathbf{r}^i_{d_m}), \mathbf{h}^i_{d_j,\text{spe}} = E_{\text{spe}}(\mathbf{r}^i_{d_j}), \mathbf{h}^i_{d_k,\text{spe}} = E_{\text{spe}}(\mathbf{r}^i_{d_k})$$

To ensure dialect-specific representations remain distinct across dialects and encode dialectal idiosyncrasies, we treat representations from different dialects as *negative* pairs and apply contrastive learning at the dialect level:

$$\mathcal{L}^{\text{dia}}_{\text{cts}} = \log \frac{1}{\sum_i f(\mathbf{h}^i_{d_m,\text{spe}}, \mathbf{h}^i_{d_j,\text{spe}}) + f(\mathbf{h}^i_{d_m,\text{spe}}, \mathbf{h}^i_{d_k,\text{spe}})}$$

Additionally, to further enforce the separation between dialect-specific and dialect-invariant features, we treat their representations within the same dialect as negative pairs and integrate their contrastive loss into the semantic loss $\mathcal{L}^{\text{sem}}_{\text{cts}}$, explicitly promoting distinction between semantic and dialectal features:

$$\mathcal{L}^{\text{sem}}_{\text{cts}} = -\log \frac{f(\mathbf{h}^i_{d_m,\text{inv}}, \mathbf{h}^i_{d_j,\text{inv}}) + f(\mathbf{h}^i_{d_m,\text{inv}}, \mathbf{h}^i_{d_k,\text{inv}})}{\sum_d \sum_{j\neq i} f(\mathbf{h}^i_{d,\text{inv}}, \mathbf{h}^j_{d,\text{inv}}) + f(\mathbf{h}^i_{d,\text{inv}}, \mathbf{h}^i_{d,\text{spe}})}$$

It should be noted that mean-pooling is applied over the frame-level features of adapter outputs to obtain fixed-dimensional utterance-level representations for obtaining the similarities and calculating the contrastive loss. With better representations from the dialect-invariant encoder, the whisper decoder could generate more accurate transcriptions. Finally, the overall loss function is the sum of the above losses, which we minimize during model training:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \alpha \mathcal{L}^{\text{sem}}_{\text{cts}} + \beta \mathcal{L}^{\text{dia}}_{\text{cts}}$$

where $\mathcal{L}_{\text{task}}$ refers to the standard ASR cross-entropy loss.

Table 12: The statistics of the audio data in SwissDial dataset.

Swiss German	AG	BE	BS	LU	GR	SG	VS	ZH	Total
Parallel Scenario									
Audio Length (hours)	2.19	2.82	2.55	1.99	2.32	3.03	2.67	2.42	19.99
Average Utterance (seconds)	3.11	4.01	3.64	2.84	3.30	4.30	3.80	3.44	3.56
Average Number of Words per Utterance	10.1	10.3	11.0	10.7	10.8	10.5	10.2	10.6	10.5

Table 13: Overall performance comparison of different models on the Swiss German dialects under parallel scenario.

	XLS-R 1B	Whisper	Dolphin	SS-kNN	XLS-R 1B-ft	Whisper-ft	Ours
WER	69.3	79.13	69.3	69.3	27.7	33.12	**17.24**
CER	33.8	30.00	33.8	33.8	13.3	12.47	**3.81**

Table 14: Overall performance comparison of different models on the Swiss German dialects under parallel scenario. Sem refers to the semantic-level contrastive learning and dia refers to the dialect-level contrastive learning.

	sg		vs		lu		gr		be		zh		ag		bs		Average	
	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer
Whisper	79.09	26.64	81.21	31.57	87.81	31.86	69.61	26.03	83.41	34.32	75.63	27.58	88.41	32.32	69.62	24.69	79.13	30.00
Whisper-ft	29.72	9.98	33.07	12.16	36.75	13.18	28.74	9.86	34.04	13.71	28.24	9.50	41.02	14.65	28.76	10.87	33.12	12.47
Ours	**18.13**	**3.68**	**18.02**	**3.71**	**21.07**	**4.8**	**16.54**	**3.3**	**18.51**	**4.34**	**16.29**	**3.15**	**25.06**	**5.72**	**16.94**	**3.96**	**17.24**	**3.81**
- w/o sem	21.82	4.72	21.41	4.55	24.94	5.94	19.76	4.24	22.91	5.06	19.48	4.13	28.47	6.61	19.68	4.89	22.73	5.76
- w/o dia	20.2	4.21	22.25	4.72	24.62	5.43	18.47	3.72	20.38	1.18	19.09	3.83	29.18	6.74	19.06	4.04	20.35	4.47

Table 15: Performance comparison for Swiss German dialects with different scales of training set under parallel scenario. **Bold** refers to the performance on the unseen dialects.

	sg		vs		lu		gr		be		zh		ag		bs		Average	
	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer
Less data																		
Whisper-ft	30.12	10.14	33.77	12.62	**73.75**	**23.18**	29.14	10.24	34.74	13.31	29.02	9.77	**81.22**	**24.38**	28.41	10.23	43.45	15.77
Ours	18.07	3.68	17.73	3.69	**56.2**	**12.12**	15.63	3.09	18.12	4.18	16.17	3.16	**63.75**	**13.82**	16.24	3.8	25.66	5.65
Mini data																		
Whisper-ft	33.24	11.13	34.21	12.73	**75.02**	**24.11**	29.33	10.73	**74.38**	**38.74**	29.61	10.02	**82.71**	**25.02**	**29.36**	**11.13**	52.44	16.17
Ours	18.12	3.73	18.54	3.76	**58.75**	**12.9**	15.36	3.03	**55.44**	**13.13**	38.71	7.69	**65.56**	**14.54**	**16.42**	**3.83**	33.77	7.56

B.3 Experiments & Results

SwissDial. The SwissDial corpus [22] constitutes a comprehensive parallel dataset encompassing eight major Swiss German dialects (AG, BE, BS, GR, LU, SG, VS, ZH). For parallel scenario, we selected only sentences spoken by speakers from all eight dialects to ensure balance across dialects. Then we randomly sampled 30% of the data to form a test set. During sampling, we ensured different recordings of the same sentence spoken by different speakers were assigned to the *same* set. This guarantees a strict separation between the training and test sets, differing from [20], and leading to a more realistic evaluation. Details of the number of utterances and hours of recordings are shown in Table 12.

The results of using the proposed pipeline under the parallel scenario are summarized in Table 14. We note significant gains over the fine-tuned baseline, reducing the WER to 17.24% and the CER to 3.81%, corresponding to absolute gains of 15.88% in WER and 8.66% in CER. Results summarized in Table 5 also show the generalizability of our proposed method.

We also perform ablation study on data augmentation under the parallel scenario. More specifically, we conducted experiments with 5 different combinations of dialects: (1) Top-2 most similar dialects; (2) Dialect most similar to the standard German (GR) and the least similar dialect; (3) Top-2 least similar dialects; (4) Dialect most similar to all the other dialects on average (ZH) and the most similar dialect; and (5) Dialect most similar to all the other dialects on average (ZH) and the least similar dialect. All similarity above is computed based on the cosine similarity of representations obtained from the Whisper model.

Table 16: Performance comparison of different augmentation methods under parallel scenario.

	sg		vs		lu		gr		be		zh		ag		bs		Average	
	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer	wer	cer
gr+top1	18.13	3.68	18.02	3.71	21.07	4.8	16.54	3.3	18.51	4.34	16.29	3.15	25.06	5.72	16.94	3.96	17.24	3.81
top2	19.08	4.83	19.76	5.02	22.01	5.87	17.77	4.47	18.60	5.04	17.47	4.33	26.39	6.99	17.40	5.02	18.46	4.94
gr+least1	21.82	4.72	21.41	4.55	24.94	5.94	19.76	4.24	20.91	5.06	19.48	4.13	28.47	6.61	19.68	4.89	22.73	5.76
least2	19.16	4.85	19.33	4.85	22.40	5.72	17.47	4.37	18.85	5.00	17.37	4.24	27.00	7.01	17.46	4.97	18.54	4.86
zh+top1	19.39	4.85	19.70	4.93	22.29	5.98	17.24	4.27	18.56	4.86	17.14	4.27	26.66	7.01	18.10	5.69	18.49	4.96
zh+least1	20.14	5.00	20.38	5.13	23.86	6.01	18.03	4.45	19.35	5.09	18.44	4.45	28.62	7.50	18.39	5.23	19.6	5.09

Table 17: Sensitivity to corruption of the retrieved cross-dialect view on SwissDial under the non-parallel setting. WER and CER are percentages.

Retrieval strategy	WER	CER
mBERT retrieval	**30.54**	**6.84**
10% random replacement	35.74	9.88
25% random replacement	40.33	12.52
50% random replacement	43.19	18.93

Table 18: Effect of replacing RINCE with InfoNCE in the semantic branch. Error rates are percentages; the remaining framework is unchanged.

Dataset	Metric	RINCE	InfoNCE
SwissDial	WER / CER	**30.54 / 6.84**	40.67 / 11.12
KeSpeech	CER	**15.82**	23.73
Casablanca	WER / CER	**53.39 / 20.23**	56.38 / 28.44

C Additional Retrieval and Loss Comparisons

The tables below give the complete retrieval-corruption and RINCE comparisons summarized in the main text. The additional training cost mainly comes from encoding three views; retrieval is offline. We provide no formal sample-complexity guarantee, and robustness in uncontrolled deployment settings remains untested. Corruption replaces the specified fraction of retrieved cross-dialect utterances with random cross-dialect samples; this probes partial pair noise rather than a complete absence of semantic correspondence.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims are stated in the Abstract and the contributions are explicitly enumerated at the end of Section 1 (Introduction). Experimental support is provided in Section 4.3 (Tables 2, 3, 4, 12, 13), which directly substantiates the WER/CER reductions claimed in the abstract.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Limitations are discussed in Section 4.2, Section 4.3.3, Section 4.3.4, and Section 5 Conclusion.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper presents an empirical framework (data augmentation + contrastive learning) and does not contain formal theorems or proofs requiring assumptions to be stated.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: Architectural and training details are provided in Section 4.2 Experimental Setups. Method formulation is fully specified in Section 3 and Appendix B. Dataset construction and splits are detailed in Section 4.1 and Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Stated in the Abstract

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.2 specifies optimizer (Adam), batch size (32), learning rate (1e-5), warmup schedule, regularization weights (= = 0.01), number of epochs (20), and number of runs averaged (5). Dataset partitioning is described in Section 4.1 and Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Section 4.2

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: Training configuration including model size (Whisper-large-v2, 24 transformer layers — Section 4.2), batch size 32 and 20-epoch cycles is specified.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The work uses publicly released dialect speech corpora and adapts publicly released pretrained models. No new human-subjects data collection was performed; the research conforms with the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: The motivation in Section 1 and Section 5 Conclusion discusses positive impact on low-resource dialect ASR and inclusion of underrepresented language varieties.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The paper releases fine-tuned dialect ASR models built on existing pretrained backbones using publicly available dialect corpora; the released artifacts do not pose elevated misuse risk beyond that of the underlying public datasets and Whisper/XLS-R models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets are credited in Section 4.1 Datasets and Section 4.2 Experimental Setups.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The Abstract commits to releasing data and code upon acceptance. The augmentation pipeline (Section 3.1) and training framework (Section 3.2; Appendix B) are documented in sufficient detail for accompanying release documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or new human-subjects data collection; all speech data are obtained from previously released, publicly available dialect corpora cited in Section 4.1.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: As stated for item 14, no new human-subjects research was conducted; the work uses only previously released public corpora, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as a component of the proposed methodology. mBERT is used only as a sentence encoder for retrieving pseudo-parallel transcripts (Section 3.1), which is a standard, non-generative use and is not an original LLM-based component requiring declaration under the NeurIPS LLM policy.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.