Accent-Unified Representations for Robust ASR via Supervised Contrastive Learning

Anonymous ACL submission

Abstract

ASR systems remain sensitive to accented speech, especially when fine-tuned on standard English data. We propose a prompt-level supervised contrastive adaptation stage that brings together utterances of the same sentence spoken by different accented speakers before downstream ASR fine-tuning. The goal is to make the encoder organize speech representations around linguistic content rather than accent-specific variation. Experiments with Wav2Vec2 and HuBERT show consistent improvements on accented English benchmarks, with the strongest gain reaching a 23.0% relative WER reduction on L2-ARCTIC. Representation analyses further show improved same-prompt alignment and more content-centered neighborhoods after adaptation.

1 Introduction

Automatic speech recognition (ASR) has made substantial progress with large self-supervised speech encoders such as wav2vec 2.0 (Baevski et al., 2020; Conneau et al., 2021), HuBERT (Hsu et al., 2021), and WavLM (Chen et al., 2022). These models learn transferable representations from large-scale unlabeled speech and provide strong initialization for downstream ASR. However, robustness to accented speech remains challenging (Hinsvark et al., 2021): ASR performance can degrade when test speakers differ from the accent distribution seen during training or fine-tuning, as highlighted by accented-speech benchmarks such as AESRC, L2-ARCTIC, and SAA (Shi et al., 2021; Zhao et al., 2018; Weinberger, 2015).

A central difficulty is that accent changes the surface realization of linguistic content. The same sentence may be realized with different phone substitutions, stress patterns, rhythm, coarticulation, and speaker-specific acoustic properties. For robust ASR, the encoder should preserve transcription-relevant linguistic content while reducing sensitivity to accent-dependent variation. Standard CTC fine-tuning (Graves et al., 2006; Baevski et al., 2020; Conneau et al., 2021) optimizes sequence prediction, but it does not explicitly constrain the representation space so that utterances with the same linguistic content remain close across speakers and accents.

Existing approaches to accented ASR often rely on supervised accented speech, data augmentation, accent-aware modeling, or specialized adaptation modules (Han et al., 2021; Shi et al., 2021; Gong et al., 2021; Li et al., 2021; Bagat et al., 2025; Turan et al., 2020; Prabhu et al., 2023; Sun et al., 2018). These methods can be effective, but they may require transcribed accented data, explicit accent labels, or additional accent-specific components. This motivates the question we address in this work: *can we improve accented ASR robustness before final ASR fine-tuning by reshaping the encoder representation space using naturally repeated linguistic content?*

We propose a prompt-level supervised contrastive adaptation stage for accented ASR. Supervised contrastive learning pulls together examples sharing the same label and separates examples with different labels (Chen et al., 2020; Khosla et al., 2020). In our setting, the label is not the speaker, accent, character, or phoneme, but the prompt identity. We exploit corpora such as L2-ARCTIC, where multiple accented speakers read the same prompts (Zhao et al., 2018): utterances of the same sentence are treated as positives across speakers and accents, while utterances with different prompts are treated as negatives.

This differs from prior supervised contrastive approaches for ASR, which define positives through augmented views for accented speech (Han et al., 2021) or at lower linguistic granularities such as character or phoneme-level units (Fu et al., 2022). By defining positives at the prompt level, our objective encourages invariance over the full utterance-

level realization of linguistic content, including pronunciation, prosody, and accent variation, while keeping the ASR architecture unchanged.

We use this idea as an intermediate adaptation step before ASR fine-tuning. The goal is simple: before training the final recognizer, we encourage the encoder to place utterances with the same linguistic content closer together, even when they are spoken by different accented speakers. We then fine-tune the adapted encoder for ASR and test whether this representation shift improves robustness to accented speech.

Our contributions are as follows:

1. We propose prompt-level supervised contrastive adaptation for accented ASR, using repeated sentences spoken by different accented speakers to align representations by linguistic content rather than accent or speaker identity.
2. We integrate this objective into a controlled three-stage pipeline that separates representation adaptation from downstream CTC fine-tuning. This allows us to isolate the effect of Stage 2 under both standard English and accented ASR fine-tuning.
3. We evaluate Wav2Vec2-Large and HuBERT-Large on clean and accented English benchmarks. We show consistent WER reductions and relate these gains to stronger same-prompt alignment and more content-centered representation spaces.

We note that the goal of this paper is not to build the strongest possible accented ASR system, but to isolate the effect of representation-level adaptation. Specifically, we test whether adding prompt-level supervised contrastive fine-tuning makes the same wav2vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021) backbones more robust than their ASR-only counterparts.

2 Related Work

Prior work on accented ASR mainly addresses robustness through supervised accented data, accent-aware adaptation, adversarial invariance, or model specialization (Sun et al., 2018; Turan et al., 2020; Li et al., 2021; Tomanek et al., 2021; Prabhu et al., 2023). AESRC introduced a large accented English benchmark and showed that accent mismatch remains a major source of ASR degradation (Shi et al., 2021). Other approaches explicitly adapt models to accent variation, for example through supervised contrastive objectives for accented ASR (Han et al., 2021) or parameter-efficient mixtures of accent-specialized LoRA experts (Bagat et al., 2025). These methods improve accented recognition, but often rely on labeled accented ASR data, accent-specific components, or explicit accent modeling.

Contrastive learning provides a complementary way to shape representation spaces. In general supervised contrastive learning, examples sharing the same label are pulled together while examples from different labels are pushed apart (Khosla et al., 2020). In speech, contrastive objectives have been used both for self-supervised pretraining and for downstream ASR robustness. Closest to our setting, prior supervised contrastive ASR methods define positives at lower linguistic levels, such as characters or phonetic units (Han et al., 2021; Fu et al., 2022). Our work differs by defining positives at the prompt level: utterances of the same sentence are treated as positives across speakers and accents. This targets utterance-level content invariance without requiring synthetic perturbations, phoneme-level annotations, or accent-specific model modules.

3 Methodology

We propose a three-stage framework, illustrated in Figure 1. Stage 1 initializes the model from a pretrained self-supervised speech encoder. Stage 2 performs prompt-level supervised contrastive adaptation on accented speech: utterances of the same sentence are pulled together across speakers and accents, while utterances with different prompts are pushed apart. An auxiliary CTC loss is added during this stage to preserve transcription-relevant information. Stage 3 discards the projection head, keeps the adapted encoder, and fine-tunes it for ASR. This design separates representation adaptation from ASR fine-tuning, making it possible to directly compare models with and without the proposed Stage 2 adaptation.

We initialize the model from an off-the-shelf pretrained speech encoder and do not perform self-supervised pretraining ourselves. The specific Wav2Vec2-Large and HuBERT-Large checkpoints used in our experiments are described in Section 4.3.

3.1 Supervised Contrastive Learning

Connectionist Temporal Classification Loss for ASR Let x_i denote an input speech utterance and

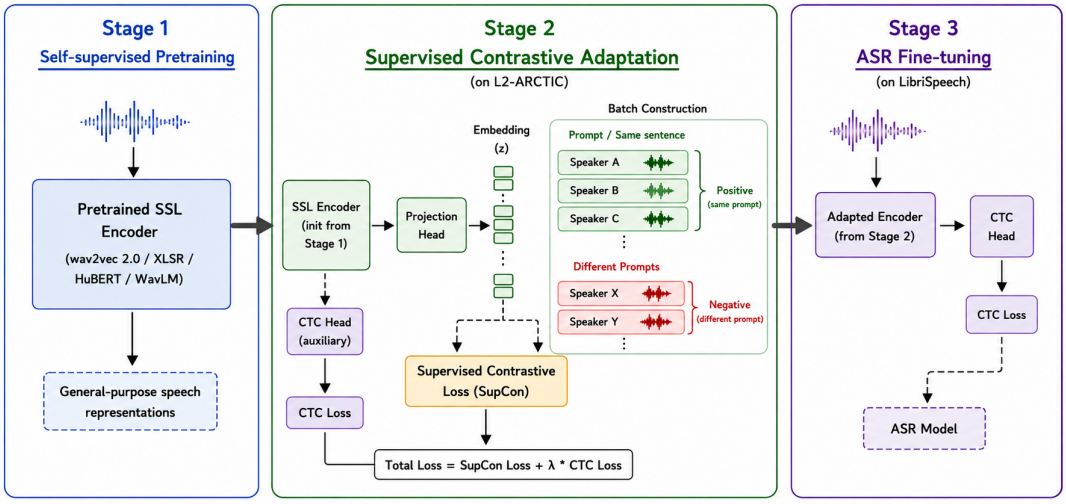

Figure 1: Overview of the proposed three-stage pipeline. Left: model initialization. Center: prompt-level supervised contrastive adaptation. Right: CTC-based ASR fine-tuning.

let $r_i = (r_{i,1}, \ldots, r_{i,U_i})$ denote its transcription, where U_i is the target sequence length. A pretrained speech encoder f_θ maps the waveform to a sequence of frame-level hidden states:

$$\mathbf{H}_i = f_\theta(x_i) = (\mathbf{h}_{i,1}, \ldots, \mathbf{h}_{i,T_i}) \in R^{T_i \times D}, \quad (1)$$

where T_i is the number of encoder time steps and D is the hidden dimension.

To preserve transcription-relevant information during contrastive adaptation, we add an auxiliary Connectionist Temporal Classification (CTC) head to the encoder (Graves et al., 2006). Given $\mathbf{H}_i$, the CTC head defines a transcription probability $p_{\text{CTC}}(r_i \mid \mathbf{H}_i)$ by marginalizing over all valid frame-level alignments between the input sequence and the reference transcription r_i. The auxiliary CTC loss over a mini-batch of size B is:

$$\mathcal{L}_{\text{CTC}} = -\frac{1}{B} \sum_{i=1}^{B} \log p_{\text{CTC}}(r_i \mid \mathbf{H}_i). \quad (2)$$

This loss acts as a transcription-preserving constraint during adaptation, preventing the contrastive objective from discarding information needed for ASR.

Supervised Contrastive Loss Since the contrastive objective is defined at the utterance level, we first convert the sequence of frame-level representations in Eq. 1 into a single utterance-level vector. We do this by applying mean pooling over the temporal dimension:

$$\bar{\mathbf{h}}_i = \frac{1}{T_i} \sum_{t=1}^{T_i} \mathbf{h}_{i,t} \in R^D, \quad (3)$$

where $\bar{\mathbf{h}}_i$ denotes the utterance-level representation.

Following contrastive learning practice (Chen et al., 2020; Khosla et al., 2020), we apply a nonlinear projection head to the utterance-level representation. The projection head is a two-layer MLP, and its output is L2-normalized before computing the contrastive loss:

$$\tilde{\mathbf{z}}_i = \text{Proj}(\bar{\mathbf{h}}_i), \qquad \mathbf{z}_i = \frac{\tilde{\mathbf{z}}_i}{\|\tilde{\mathbf{z}}_i\|_2} \in R^d, \quad (4)$$

where $d < D$ is the dimension of the contrastive space. The normalized vector $\mathbf{z}_i$ is used only for the supervised contrastive objective.

Each mini-batch is constructed by sampling K utterance prompts and S speakers per prompt. The batch size is therefore:

$$B = K \times S. \quad (5)$$

Let $\mathcal{B}$ denote the resulting mini-batch. For an anchor utterance $i \in \mathcal{B}$, we denote its linguistic content by u_i, corresponding to the prompt or sentence being spoken. The positive set for i contains all other utterances in the batch with the same linguistic content:

$$P(i) = \{p \in \mathcal{B} \setminus \{i\} \mid u_p = u_i\}. \quad (6)$$

All utterances with different linguistic content are treated as negatives, including utterances spoken by the same speaker if their prompt differs. This encourages the model to organize the contrastive space according to linguistic content rather than speaker identity or accent.

Using these positive sets, the supervised contrastive objective is defined as:

$$\mathcal{L}_{\text{supcon}} = \sum_{i \in \mathcal{B}} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{Z_i}, \tag{7}$$

where the contrastive normalization term Z_i is defined as:

$$Z_i = \sum_{a \in \mathcal{B} \setminus \{i\}} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau). \tag{8}$$

Here, $\mathbf{z}_i \cdot \mathbf{z}_a$ denotes cosine similarity between L2-normalized embeddings, and τ is the temperature parameter.

Combined together, the supervised contrastive objective and the auxiliary ASR objective define the Stage 2 adaptation loss:

$$\mathcal{L}_{\text{Stage2}} = \mathcal{L}_{\text{supcon}} + \lambda \mathcal{L}_{\text{CTC}}, \tag{9}$$

where λ controls the weight of the auxiliary CTC loss. The CTC term keeps the adaptation oriented toward ASR by preserving transcription-relevant information.

3.2 ASR Fine-tuning

After supervised contrastive adaptation, we discard the projection head and retain only the adapted speech encoder. We then follow the standard fine-tuning paradigm (Baevski et al., 2020; Conneau et al., 2021): a new CTC prediction head is added on top of the encoder, and the resulting ASR model is fine-tuned on paired speech–text data.

In our main setting, this stage is performed on regular L1 English speech. This allows us to test whether contrastive adaptation on accented speech improves downstream robustness without requiring large amounts of supervised accented ASR data. The model is trained with the standard CTC objective:

$$\mathcal{L}_{\text{Stage3}} = \mathcal{L}_{\text{CTC}}. \tag{10}$$

This stage follows the standard SSL-to-ASR transfer setting, but starts from an encoder that has first been adapted to reduce accent-related variation.

4 Experimental Setup

4.1 Datasets

L2-ARCTIC We use L2-ARCTIC (Zhao et al., 2018) as the corpus for Stage 2 supervised contrastive adaptation and for accented-speech evaluation. It contains non-native English speech from 24 speakers covering six first-language (L1) groups: Arabic, Chinese, Hindi, Korean, Spanish, and Vietnamese, with four speakers per L1 group. Speakers read CMU ARCTIC prompts, making the corpus suitable for prompt-level supervised contrastive learning because the same linguistic content is available across multiple speakers and accents.

LibriSpeech We use LibriSpeech train-clean-100 (Panayotov et al., 2015) for Stage 3 ASR fine-tuning in the main setting. LibriSpeech provides clean L1 English speech paired with transcripts. We evaluate on LibriSpeech test-clean to measure whether the proposed adaptation preserves performance on non-accented English speech.

AESRC We use AESRC (Shi et al., 2021) as an additional accented English corpus. The corpus covers 10 accents, with two accents reserved for testing. We use it either for accented ASR fine-tuning (Stage 3) or as an external accented evaluation benchmark, depending on the experimental condition.

Speech Accent Archive We use the Speech Accent Archive (SAA) (Weinberger, 2015) as an external accented-speech evaluation corpus. It contains speakers reading the same elicitation paragraph, providing a complementary benchmark for evaluating robustness to diverse English accents.

Data splits For L2-ARCTIC, we use a fixed evaluation split. We hold out one speaker from each of the six L1 groups for testing, and use the three remaining speakers for Stage 2 training and development. We also split the prompt list into non-overlapping train, development, and test prompts, using approximately 80%, 10%, and 10% of prompts, respectively. As a result, L2-ARCTIC test utterances come from speakers and prompts that are never seen during Stage 2 adaptation. All L2-ARCTIC results are reported on this fixed split.

For LibriSpeech, we use the official train-clean-100 and test-clean splits. These splits are speaker-disjoint by construction, following the original corpus design (Panayotov et al., 2015). For AESRC, when it is used for ASR fine-tuning, we reserve 10% of the training set as a held-out development split. This development split is kept separate and is not included in the training set. All random splits and training runs use a fixed random seed of 42.

Table 2 in Appendix A reports the main statistics of the corpora used for training, adaptation, and evaluation.

4.2 Experimental Conditions

We evaluate four experimental conditions designed to isolate the contribution of the proposed contrastive adaptation stage and the effect of supervised accented ASR data. The conditions differ in whether Stage 2 supervised contrastive adaptation is applied, and in whether Stage 3 ASR fine-tuning uses only standard English speech or accented English speech.

Condition A: proposed setting. The model first undergoes Stage 2 supervised contrastive adaptation on L2-ARCTIC, then Stage 3 ASR fine-tuning on LibriSpeech train-clean-100. This condition tests whether contrastive adaptation alone improves accented ASR robustness when the final ASR training data contains only standard L1 English speech.

Condition B: accented-data upper bound. The model uses the same Stage 2 supervised contrastive adaptation as Condition A, but Stage 3 ASR fine-tuning is performed on AESRC. This condition measures how much additional benefit is obtained when accented supervised ASR data is available after contrastive adaptation.

Condition C: ASR-only baseline. The model skips Stage 2 and is directly fine-tuned for ASR on LibriSpeech train-clean-100. This is the standard SSL-to-ASR transfer baseline and isolates the effect of the proposed contrastive adaptation by comparison with Condition A.

Condition D: accented ASR oracle. The model skips Stage 2 and is directly fine-tuned for ASR on AESRC. This condition provides an in-domain accented ASR reference without contrastive adaptation.

The comparison between Conditions A and C isolates the contribution of Stage 2 supervised contrastive adaptation. The comparison between Conditions A and B estimates the residual benefit of supervised accented ASR data after adaptation. Finally, Conditions C and D quantify the gap between standard English ASR fine-tuning and in-domain accented ASR fine-tuning without contrastive adaptation.

4.3 Training Setup

Models To study whether supervised contrastive adaptation improves robustness to accented speech across different self-supervised speech encoders, we use two widely used large pretrained ASR models: Wav2Vec2-Large[1] (Baevski et al., 2020) ($L = 24, H = 1024, A = 16, \sim 317$M parameters) and HuBERT-Large[2] (Hsu et al., 2021) ($L = 24, H = 1024, A = 16, \sim 316$M parameters). The two models have comparable scale but rely on different self-supervised pretraining objectives: wav2vec 2.0 (w2v2) uses contrastive prediction over masked latent speech representations, whereas HuBERT uses offline clustering to define discrete acoustic-unit targets and learns to predict them in masked regions. This allows us to test whether prompt-level supervised contrastive adaptation transfers across different speech representation-learning paradigms.

Training details For Stage 2, each mini-batch is constructed by sampling $K = 15$ prompts and $S = 12$ speakers per prompt, giving an effective batch size of 180 utterances. The projection head is a two-layer MLP with hidden dimension 512 and output dimension 256. We set the supervised contrastive temperature to $\tau = 0.1$ and the auxiliary CTC weight to $\lambda = 0.1$. Stage 2 is trained for 50 epochs. We use AdamW with learning rate 2×10^{-5}, weight decay 10^{-4}, gradient clipping at 1.0, and 500 warmup steps.

For Stage 3, the projection head is discarded and a new CTC head is initialized for ASR fine-tuning. Conditions A and C are fine-tuned on LibriSpeech train-clean-100, while Conditions B and D are fine-tuned on the AESRC training split. In our AESRC setup, the split contains 123,531 training utterances and 12,998 development utterances. All Stage 3 systems are trained for up to 30k optimization steps with batch size 16, encoder learning rate 10^{-5}, CTC-head learning rate 10^{-4}, weight decay 10^{-2}, warmup ratio 0.1, and gradient clipping at 1.0. The best checkpoint is selected according to development-set WER. Full training hyperparameters and decoding details are reported in Appendix B and Appendix C, respectively.

5 Results

5.1 Downstream ASR Performance

Table 1 reports downstream ASR performance in WER on LibriSpeech test-clean, L2-ARCTIC, AESRC, and SAA. Lower values indicate better recognition performance. We focus on three controlled comparisons: C–A, D–B, and A–B.

[1] `facebook/wav2vec2-large-960h`
[2] `facebook/hubert-large-ls960-ft`

Table 1: Downstream ASR performance in WER (%). Lower is better. For each backbone separately, bold numbers indicate the best score among its corresponding systems. C–A isolates the effect of Stage 2 with LibriSpeech ASR fine-tuning, while D–B evaluates the same effect with AESRC ASR fine-tuning. Whisper-Large-v3 (Radford et al., 2023) is reported as an external reference.

Model	Condition	LibriSpeech	L2-ARCTIC	AESRC	SAA
Wav2Vec2-Large	HF checkpoint	**4.31**	26.06	23.51	22.14
	C: LibriSpeech	5.24	24.20	24.59	19.65
	A: SupCon + LibriSpeech	4.62	**18.63**	21.61	**17.84**
	D: AESRC	6.65	24.08	18.46	18.52
	B: SupCon + AESRC	5.49	23.89	**17.38**	18.12
HuBERT-Large	HF checkpoint	**2.11**	16.90	16.25	11.54
	C: LibriSpeech	2.18	17.94	16.60	11.69
	A: SupCon + LibriSpeech	2.18	15.79	15.76	**11.48**
	D: AESRC	5.70	15.66	11.96	13.59
	B: SupCon + AESRC	5.10	**15.58**	**11.71**	13.20
Whisper-Large-v3	External reference	2.37	11.20	12.61	9.90

SupCon improves robustness without accented ASR data (A vs. C). For Wav2Vec2-Large, Stage 2 gives consistent improvements over the ASR-only baseline. On the accented benchmarks, WER decreases from 24.20 to 18.63 on L2-ARCTIC, corresponding to a **23.0% relative reduction**; from 24.59 to 21.61 on AESRC, a **12.1% relative reduction**; and from 19.65 to 17.84 on SAA, a **9.2% relative reduction**. LibriSpeech WER also improves from 5.24 to 4.62, showing that the gain in accented robustness does not come at the expense of clean-speech ASR in this setting. Compared with the original HF checkpoint, however, the final LibriSpeech WER remains slightly higher, 4.62 vs. 4.31. Thus, the main benefit of Stage 2 is observed on accented evaluation sets, while the cost on standard clean English remains small.

For HuBERT-Large, the same trend holds, but with smaller gains. On L2-ARCTIC, WER decreases from 17.94 to 15.79, corresponding to a **12.0% relative reduction**. On AESRC, WER decreases from 16.60 to 15.76, a **5.1% relative reduction**. On SAA, the improvement is modest, from 11.69 to 11.48. LibriSpeech performance remains unchanged between C and A at 2.18 WER. Compared with the original HF checkpoint, this corresponds to a small degradation on clean speech, from 2.11 to 2.18, and SAA also remains slightly worse than the HF checkpoint. Nevertheless, Condition A obtains the best HuBERT results on L2-ARCTIC and AESRC, suggesting that Stage 2 improves the accented benchmarks most directly aligned with our robustness objective, while preserving clean ASR performance within a small margin.

SupCon remains useful with accented ASR data (B vs. D). The B–D comparison tests whether Stage 2 remains useful when Stage 3 already uses supervised accented speech from AESRC. For Wav2Vec2-Large, adding Stage 2 improves all evaluation sets. The largest gain is obtained on AESRC, where WER decreases from 18.46 to 17.38, corresponding to a **5.9% relative reduction**. Improvements on the other sets are smaller: LibriSpeech improves from 6.65 to 5.49, L2-ARCTIC from 24.08 to 23.89, and SAA from 18.52 to 18.12.

HuBERT-Large shows the same overall pattern. Condition B improves over Condition D on all evaluation sets, with WER decreasing from 5.70 to 5.10 on LibriSpeech, from 15.66 to 15.58 on L2-ARCTIC, from 11.96 to 11.71 on AESRC, and from 13.59 to 13.20 on SAA. These gains are modest but consistent. This suggests that Stage 2 is not redundant with accented ASR fine-tuning, although its additional benefit is smaller than in the A–C setting because AESRC fine-tuning already captures part of the accent-robustness benefit.

AESRC fine-tuning favors in-domain performance over balance (A vs. B). The A–B comparison measures the effect of replacing LibriSpeech with AESRC during Stage 3, after the same Stage 2 contrastive adaptation. Overall, AESRC fine-tuning improves performance on the matched AESRC test set, but does not consistently improve cross-corpus robustness.

For Wav2Vec2-Large, AESRC WER decreases

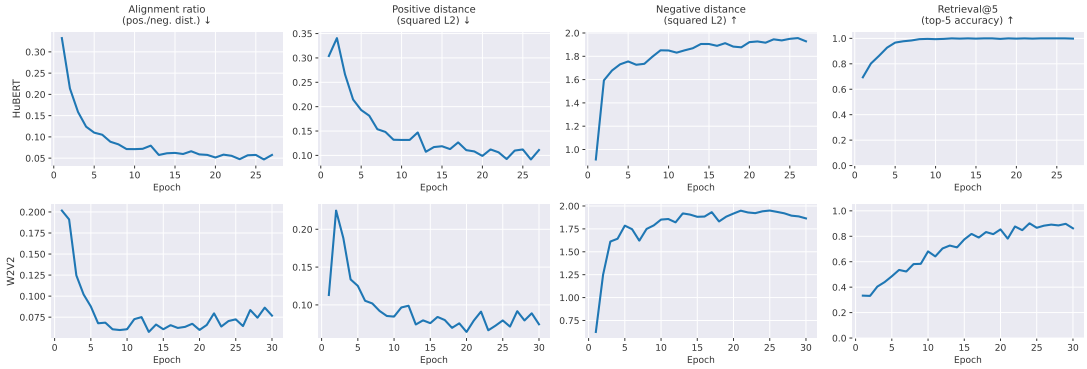

Figure 2: Representation-space alignment during Stage 2 for HuBERT and Wav2Vec2. Positive and negative distances are squared L2 distances between same-prompt and different-prompt utterances, respectively. Lower alignment ratio and higher Retrieval@5 indicate stronger prompt-level organization.

from 21.61 to 17.38, corresponding to a **19.6% relative reduction**. However, this specialization comes with worse performance on the other evaluation sets: LibriSpeech WER increases from 4.62 to 5.49, L2-ARCTIC from 18.63 to 23.89, and SAA from 17.84 to 18.12. Thus, AESRC fine-tuning improves in-domain accented recognition, but reduces the more balanced generalization obtained with Condition A.

For HuBERT-Large, the pattern is similar but slightly less uniform. AESRC fine-tuning strongly improves AESRC WER from 15.76 to 11.71, a **25.7% relative reduction**, and slightly improves L2-ARCTIC from 15.79 to 15.58. However, it substantially degrades LibriSpeech performance from 2.18 to 5.10 and worsens SAA from 11.48 to 13.20. Overall, Condition B provides stronger AESRC specialization, while Condition A remains the more balanced setting across clean and external accented evaluation sets.

5.2 SupCon Representation Dynamics

We analyze whether Stage 2 changes the representation space in the intended direction. Figure 2 tracks four metrics during adaptation: positive distance between same-prompt utterances, negative distance between different-prompt utterances, their alignment ratio, and retrieval@5. Lower positive distance, higher negative distance, lower alignment ratio, and higher retrieval@5 all indicate stronger prompt-level organization.

For HuBERT, the effect is rapid and stable. The alignment ratio drops sharply during the first epochs, from roughly 0.33 to below 0.10, and then remains low. This is driven by both decreasing positive distances and increasing negative distances. Retrieval@5 quickly approaches 1.0, showing that same-prompt utterances become almost perfectly recoverable from local neighborhoods. This indicates that HuBERT rapidly forms a highly structured content-centered space.

For Wav2Vec2, the same trend appears, but more slowly and with noisier dynamics. The alignment ratio decreases early in training, positive distances decrease overall, and negative distances increase and remain high. Retrieval@5 improves progressively from around 0.3 to nearly 0.9 by the end of training. Thus, Wav2Vec2 also learns prompt-level organization, but requires more epochs than HuBERT.

Overall, Stage 2 produces the intended geometric effect: same-prompt utterances move closer together, different-prompt utterances move farther apart, and same-prompt retrieval improves substantially. This supports the hypothesis that prompt-level SupCon adaptation organizes utterance-level representations around linguistic content rather than speaker- or accent-specific variation.

5.3 t-SNE Visualization of Final ASR Representations

To complement the quantitative alignment metrics, we visualize the final Stage 3 backbone representations using t-SNE. Figure 3 compares Condition C, where Wav2Vec2-Large is directly fine-tuned for ASR on LibriSpeech, with Condition A, where the model first undergoes Stage 2 supervised contrastive adaptation. The same L2-ARCTIC subset is used in all panels. Colors indicate either utterance identity, corresponding to prompt-level content, or speaker accent.

Condition C shows a relatively diffuse struc-

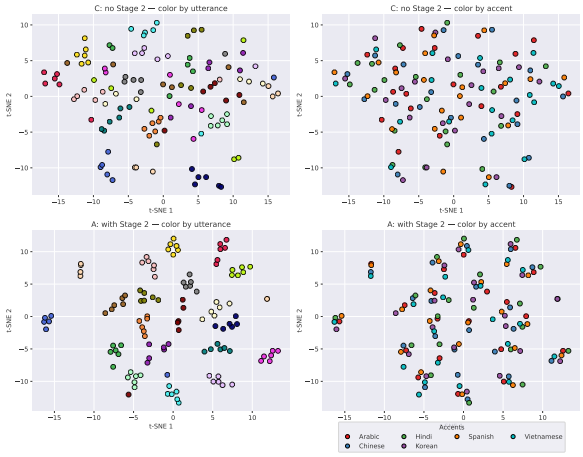

Figure 3: t-SNE visualization of Wav2Vec2 Stage 3 backbone representations on L2-ARCTIC. Condition C skips Stage 2, while Condition A uses supervised contrastive adaptation before ASR fine-tuning. Left panels are colored by utterance identity, corresponding to prompt-level linguistic content; right panels show the same points colored by accent. In Condition A, same-prompt utterances form compact clusters containing multiple accents, indicating that Stage 2 encourages content-centered rather than accent-dominated representations. Utterance labels and transcripts are listed in Appendix E.

ture: utterances corresponding to the same prompt are not consistently grouped into compact neighborhoods. After Stage 2 adaptation, the geometry changes substantially. In Condition A, same-prompt utterances form clear local clusters, indicating stronger organization by linguistic content. Importantly, the accent-colored view shows that these clusters contain multiple accents rather than a single dominant accent. This supports the intended effect of the proposed objective: utterances with the same linguistic content are brought together despite accent variation.

This visualization should be interpreted as qualitative evidence rather than standalone proof, since t-SNE is a nonlinear projection. Nevertheless, it is consistent with Figure 2: positive distances decrease, negative distances increase, and retrieval@5 improves during Stage 2. Together, these results suggest that supervised contrastive adaptation reshapes the embedding space toward content-centered neighborhoods that are less dominated by accent-specific variation.

We observe the same qualitative trend for HuBERT-Large, reported in Appendix D. Compared with Wav2Vec2, HuBERT starts from a more structured representation space before Stage 2, which may partly explain the smaller downstream gains in Table 1. This is consistent with HuBERT's masked-prediction pretraining objective, which predicts discrete hidden-unit targets and may already impose stronger content-related structure. Stage 2 still improves the organization of the space, but the visual and downstream effects are less pronounced than for Wav2Vec2.

6 Conclusion

We introduced prompt-level supervised contrastive adaptation for accented ASR. The method uses repeated sentences spoken by different accented speakers to bring same-content utterances closer in the representation space before downstream ASR fine-tuning. Across Wav2Vec2-Large and HuBERT-Large, this adaptation improves accented ASR performance, with the clearest gains when the final ASR model is fine-tuned on LibriSpeech.

Representation analyses support this interpretation: Stage 2 reduces same-prompt distances, improves same-prompt retrieval, and produces more content-centered clusters. When ASR fine-tuning uses AESRC, the gains are smaller but still consistent, suggesting that prompt-level adaptation remains complementary to supervised accented ASR data. Overall, these results show that reshaping the encoder representation space before ASR fine-tuning is a simple and effective way to improve robustness to accented speech.

7 Limitations and Future Work

This work has several limitations. First, the proposed adaptation relies on corpora where multiple speakers read the same prompts, as in L2-ARCTIC. Extending it to less controlled domains or spontaneous speech would require alternative positive-pair definitions, such as transcript matching or semantic similarity.

Second, our L2-ARCTIC results are reported on one fixed speaker- and prompt-disjoint split rather than averaged over multiple folds. Although we also evaluate on external accented benchmarks such as AESRC and SAA, future work should report multi-fold results to better quantify variance.

Finally, Whisper-Large-v3 remains stronger on several evaluation sets, suggesting that large encoder–decoder ASR systems still provide stronger robustness. Future work should extend prompt-level contrastive adaptation to stronger generative ASR architectures.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*.

Raphaël Bagat, Irina Illina, and Emmanuel Vincent. 2025. Mixture of lora experts for low-resourced multi-accent automatic speech recognition. In *Proceedings of Interspeech 2025*, pages 1143–1147. Also available as arXiv:2505.20006.

Sanyuan Chen, 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 Zeng, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech*.

Li Fu, Xiaoxiao Li, Runyu Wang, Lu Fan, Zhengchen Zhang, Meng Chen, Youzheng Wu, and Xiaodong He. 2022. SCaLa: Supervised contrastive learning for end-to-end speech recognition. In *Proceedings of Interspeech*, pages 2598–2602.

Xun Gong, Yizhou Lu, Zhikai Zhou, and Yanmin Qian. 2021. Layer-wise fast adaptation for end-to-end multi-accent speech recognition. In *Proceedings of Interspeech*, pages 1274–1278.

Alex Graves, 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*, pages 369–376.

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

Arthur Hinsvark, Natalie Delworth, Miguel Del Rio, Quinten McNamara, Joshua Dong, Ryan Westerman, Michelle Huang, Joseph Palakapilly, Jennifer Drexler, Ilya Pirkin, Nishchal Bhandari, and Miguel Jette. 2021. Accented speech recognition: A survey. *arXiv preprint arXiv:2104.10747*.

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

Jitsi. 2025. JiWER: Similarity measures for automatic speech recognition evaluation. `https://github.com/jitsi/jiwer`. Python package.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673.

Jialu Li, Vimal Manohar, Pooja Chitkara, Andros Tjandra, Michael Picheny, Frank Zhang, Xiaohui Zhang, and Yatharth Saraf. 2021. Accent-robust automatic speech recognition using supervised and unsupervised wav2vec embeddings. *Preprint*, arXiv:2110.03520.

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

Darshan Prabhu, Preethi Jyothi, and Florian Metze. 2023. Accented speech recognition with accent-specific codebooks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7175–7188.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Xian Shi and 1 others. 2021. The accented english speech recognition challenge 2020: Open datasets, tracks, baselines, results and methods. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

Sining Sun, Ching-Feng Zhang, and Philip C. Woodland. 2018. Domain adversarial training for accented speech recognition. *arXiv preprint arXiv:1806.02786*.

Katrin Tomanek, Vicky Zayats, Dirk Padfield, Kara Vaillancourt, and Fadi Biadsy. 2021. Residual adapters for parameter-efficient asr adaptation to atypical and accented speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6751–6760.

Mehmet Ali Turan, Emmanuel Vincent, and Denis Jouvet. 2020. Achieving multi-accent asr via unsupervised acoustic model adaptation. In *Proceedings of Interspeech*, pages 1286–1290.

Steven Weinberger. 2015. Speech accent archive. George Mason University. Available at https://accent.gmu.edu/.

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

A Dataset Statistics

Table 2: Summary statistics of the corpora used for training and evaluation.

Dataset	Speakers	Utterances	Size
L2-ARCTIC	24	26,867	27.1h
AESRC	∼600	123,531	200h
LibriSpeech train	251	28,539	100h
LibriSpeech test-clean	40	2,620	5.4h
SAA	3,038	3,038	∼6h

B Training Hyperparameters

Stage	Hyperparameter	Value
Stage 2	Batch construction	$K = 15$ prompts, $S = 12$ speakers per prompt
	Effective batch size	180 utterances
	Projection head	2-layer MLP
	Projection hidden dimension	512
	Projection output dimension	256
	SupCon temperature	$\tau = 0.1$
	Auxiliary CTC weight	$\lambda = 0.1$
	Optimizer	AdamW
	Learning rate	2×10^{-5}
	Weight decay	10^{-4}
	Gradient clipping	1.0
	Warmup steps	500
	Training length	50 epochs
Stage 3	ASR objective	CTC
	Encoder learning rate	10^{-5}
	CTC-head learning rate	10^{-4}
	Weight decay	10^{-2}
	Warmup ratio	0.1
	Gradient clipping	1.0
	Batch size	16
	Maximum optimization steps	30k
	Checkpoint selection	Best development-set WER

Table 3: Hyperparameters used for Stage 2 supervised contrastive adaptation and Stage 3 ASR fine-tuning.

C Decoding and Evaluation

We use greedy CTC decoding for all ASR systems, without beam search, language model, or lexicon constraints. Hypotheses and references are lowercased and stripped of punctuation before scoring. WER is computed with `jiwer` (Jitsi, 2025).

D HuBERT t-SNE Visualization

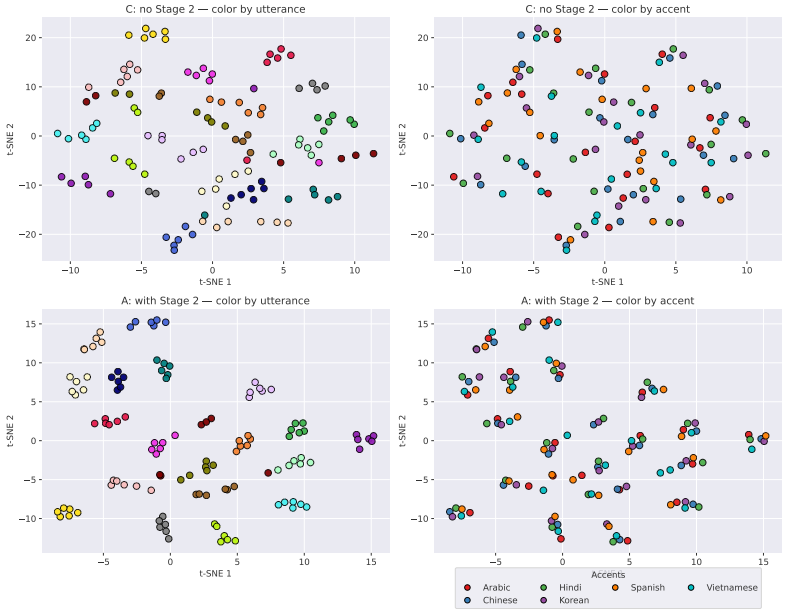

Figure 4: Additional t-SNE visualization for HuBERT-Large using the same analysis protocol as Figure 3.

E Utterance Mapping for t-SNE Visualizations

ut_id	transcript	color
UT1	i was the only one who remained sitting	(#e6194b)
UT2	how could he explain his possession of the sketch	(#3cb44b)
UT3	he looked at the handkerchief more closely	(#ffe119)
UT4	shall i carry you	(#4363d8)
UT5	she was sleeping under his protection as sweetly as a child	(#f58231)
UT6	her words sent a strange chill through philip	(#911eb4)
UT7	two years ago i gave up civilization for this	(#46f0f0)
UT8	about him everywhere were the evidences of luxury and of age	(#f032e6)
UT9	the ringing of the big bell aroused him	(#bcf60c)
UT10	she nodded and her eyes grew soft and moist	(#fabebe)
UT11	i never allow what can't be changed to annoy me	(#008080)
UT12	think of investing in such an adventure	(#e6beff)
UT13	daylight found himself charmed and made curious by the little man	(#9a6324)
UT14	and right there i saw and knew it all	(#fffac8)
UT15	he was an enthusiast and a desert dweller	(#800000)
UT16	he had proved it today with his amateurish and sophomoric productions	(#aaffc3)
UT17	gregson had seated himself under the lamp and was sharpening a pencil	(#808000)
UT18	i'll only be in the way	(#ffd8b1)
UT19	i know that you are in charge there and jeanne knows	(#000075)
UT20	the boy threw back his head with pride	(#808080)

Table 4: Utterance transcripts and associated colors used in the t-SNE visualizations.