Multimodal Consistency-Guided Reference-Free Data Selection for ASR Accent Adaptation

Ligong Lei, Wenwen Lu, Xudong Pang, Zaokere Kadeer, Aishan Wumaier*
School of Computer Science and Technology, Xinjiang University, Urumqi, China
Xinjiang Key Laboratory of Multilingual Information Technology, Urumqi, China
Joint International Research Laboratory of Silk Road Multilingual Cognitive Computing, Urumqi, China
{107552304075, 107552301358, 107552304120}@stu.xju.edu.cn, {zuhra, Hasan1479}@xju.edu.cn

***Abstract*—Automatic speech recognition (ASR) systems often degrade on accented speech because acoustic-phonetic and prosodic shifts induce a mismatch to training data, making labeled accent adaptation costly. Text-centric pseudo-label filters (e.g., perplexity) can favor fluent hypotheses that remain acoustically misaligned; using them as supervision during fine-tuning amplifies errors. We present a reference-free selection pipeline for accent adaptation under a transductive, label-free protocol: unlabeled target speech drives pool reduction and percentile thresholds without oracle transcripts. We optionally apply target-aware FLMI preselection, then decode several hypotheses per utterance and score them with speech–text alignment in a frozen SONAR embedding space and predicted word error rate (WER), treating multimodal consistency as a proxy for pseudo-label quality under accent-induced mismatch. A simple percentile rule selects the training subset. In-domain, selecting ∼1.5k utterances from a 30k pool achieves 10.91% WER, close to 10.45% with 30k supervised labels. With a mismatched cross-domain candidate pool, consistency-filtered subsets avoid degradation from unfiltered pseudo-labels under strong accent shift; matched-hour experiments on a stronger ASR backbone further confirm gains over random sampling and recent selection baselines.**

***Index Terms*—Data selection, speech recognition, accent adaptation, multimodal models, word error rate prediction**

I. Introduction

Automatic Speech Recognition (ASR) enables many artificial intelligence applications, including voice assistants, captioning, call analytics, and accessibility tools. Recent progress in end-to-end modeling and self-supervised learning (SSL) pre-training has improved accuracy and robustness, yet ASR performance still drops substantially on accented speech. This failure is largely driven by accent-induced shifts in acoustic-phonetic realization and prosody that create a mismatch to training data. In practice, deploying high-performance accent-adapted ASR systems often requires substantial labeled data and compute, which increases cost and slows iteration.

Data selection provides a complementary direction for improving training efficiency on large-scale corpora by identifying and retaining the most informative samples. By removing redundant or noisy data, selection aims to reduce compute, time, and financial burden while maintaining, or even improving, generalization [1].

In accent adaptation, pseudo-labeling unlabeled target-domain speech is particularly attractive. However, selecting reliable pseudo-labels is non-trivial because accent mismatch primarily manifests in acoustics and prosody rather than in semantics. Text-centric heuristics (e.g., perplexity filtering) can therefore rank fluent transcripts that remain acoustically mismatched; using them as supervision amplifies errors during fine-tuning. In contrast, while acoustic realizations shift under accents, the underlying semantic content is typically preserved, which motivates using cross-modal consistency as a reference-free proxy for pseudo-label quality. Prior work on accent adaptation typically mitigates mismatch by fine-tuning a strong speaker-independent recognizer using a small amount of accent-specific labeled data [2], [3], or by prioritizing sentences that are likely to trigger ASR errors for targeted recording and personalization [4]. These strategies, however, generally assume access to annotated target data. In this work, we instead study a pseudo-labeling setting where the goal is to identify reliable training utterances from a candidate pool without ground-truth transcripts.

We ask how reference-free signals can indicate pseudo-

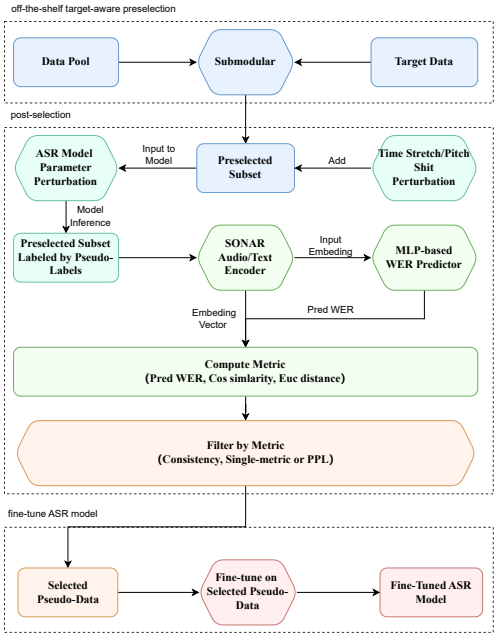

Fig. 1. Overall pipeline of our unsupervised data-selection framework.

*Corresponding authors.

label quality when mismatch is primarily acoustic. Following target-aware FLMI preselection [5] as an optional pool-reduction step, we operationalize *multimodal consistency*: frozen SONAR speech–text alignment [6] and reference-free WER prediction score each hypothesis in a perturbation set, so cross-modal agreement acts as a proxy for whether a pseudo-label tracks the utterance under accent shift. A simple percentile acceptance rule yields a label-free training subset. We frame this as a systems-level pipeline—optional FLMI, perturbation decoding, consistency scoring, and thresholding—rather than a new ASR objective. Our contributions are threefold:

- We assemble a label-free, transductive selection procedure for accent adaptation by chaining optional FLMI-based pool reduction, perturbation decoding, multimodal consistency scoring, and a percentile retention rule—reusing established building blocks where appropriate and focusing evaluation on the full protocol.
- Within this protocol, we use frozen SONAR speech–text alignment and reference-free WER prediction to score competing hypotheses under accent-induced mismatch, where text-only perplexity is a weak proxy.
- We empirically validate the integrated pipeline across in-domain and cross-domain pools, backbones, and matched-hour budgets, including settings where unfiltered pseudo-labels degrade performance.

II. Related Work

Data Selection. SSL-embedding-similarity-based selection measures self-supervised learning (SSL) embedding similarity to select data from a labeled source pool and boosts cross-domain ASR performance [7]. However, this line of work focuses on labeled datasets and does not consider pseudo-label selection from unlabeled pools. More practical pipelines first automatically transcribe unlabeled speech and then filter for valuable segments. Perplexity-contrast filtering retains domain-relevant pseudo-labeled utterances by contrasting the perplexity of a general language model against that of an in-domain language model [8]. A domain-classifier-based selection pipeline employs Whisper-X to obtain pseudo-transcriptions, uses TF-IDF and a support vector machine (SVM) for domain classification, and filters pseudo-labels based on word ratio and perplexity [9]. While these methods are effective for domain adaptation, their text-centric heuristics often underrepresent acoustic-phonetic and prosodic factors central to accented speech, thus missing accent-induced acoustic discrepancies. DITTO uses submodular mutual information on Mel-frequency cepstral coefficient (MFCC) features to obtain fair accent-specific subsets, primarily selecting data from an acoustic perspective, but relies on human-annotated transcripts rather than pseudo-labels [10]. In a more recent study, WER-based pseudo-label filtering applies an SVM binary classifier with SSL features (0.5 threshold) and combines named entity recognition (NER) and character error rate (CER) for data selection [11]; although this approach considers the match between acoustic signals and pseudo-labels through WER prediction, the selection granularity remains coarse, relying primarily on binary accept/reject decisions rather than continuous, multi-factor quality scoring.

Pseudo-label Quality Estimation. Reference-free WER estimation sees rapid progress in recent years. A reference-less ASR quality metric leverages known quality relationships between multiple compressed outputs of an ASR system, fine-tuned in a self-supervised manner using contrastive learning and a Siamese network architecture to pairwise rank ASR hypotheses [12]. Meanwhile, FeWER [13] utilizes SSL features and a lightweight multilayer perceptron (MLP) to predict sentence-level WER; its system-agnostic variant further eliminates dependence on a specific decoder through a hypothesis generator [14]. These advances lay a solid foundation for pseudo-label-based selection by providing reliable, reference-free quality signals. Our approach does not use WER prediction in isolation, but integrates it with cross-modal alignment scores to construct a unified consistency metric, which is then applied to data selection.

Multimodal Representation. Sentence-level representation learning includes encoder-only, encoder-decoder, and teacher-student frameworks. SONAR [6] employs a two-stage design: an encoder-decoder model first builds a multilingual text embedding space, after which speech encoders are trained in a teacher-student manner to align speech into the same space, enabling cross-lingual and cross-modal similarity search and translation mining. In our work, we utilize frozen SONAR speech/text encoders to obtain aligned sequence-level embeddings. This shared space enables measuring the semantic fidelity between speech and its pseudo-transcription, which in turn supports cross-modal, reference-free quality estimation within our selection pipeline.

Despite these advances, few works address *label-free* pseudo-label selection for accent adaptation with quality signals that reflect speech–text consistency under accent-induced acoustic shifts. We propose a transductive, reference-free selection framework that leverages multimodal consistency signals, detailed next.

III. Method

Fig. 1 illustrates the data flow of our selection framework. Given a large candidate pool $\mathcal{C}$ and an unlabeled target query set $\mathcal{T}$, we (i) optionally apply a target-aware FLMI preselection (prior work) to obtain a reduced pool $\mathcal{P} \subseteq \mathcal{C}$; (ii) for each utterance $u \in \mathcal{P}$, decode a baseline hypothesis H_0 and a set of perturbed hypotheses $\{H_k\}$; (iii) score each hypothesis with reference-free signals—predicted WER and speech–text alignment in the shared SONAR embedding space; and (iv) apply the acceptance rule in Sec. III-D to keep at most one hypothesis per utterance, yielding the final selected training subset $\mathcal{S}$ used for fine-tuning. Importantly, our contribution centers on steps (ii)–(iv), which can also be applied on a fixed candidate pool without FLMI.

A. FLMI Preselection via Submodular Mutual Information

We use FLMI as a target-aware preselection step following prior work [5], [10]. This stage is not our main contribution;

rather, it provides query relevance and reduces the computational burden of the subsequent perturbation-and-consistency filtering. We briefly review the necessary definitions for completeness. A set function $f : 2^{\mathcal{U}} \to \mathbb{R}$ defined on a ground set $\mathcal{U}$ of n data points is called *submodular* [15] if it satisfies diminishing marginal returns: for all $\mathcal{X} \subseteq \mathcal{Y} \subseteq \mathcal{U}$ and $j \notin \mathcal{Y}$,

$$f(j \mid \mathcal{X}) = f(\mathcal{X} \cup \{j\}) - f(\mathcal{X}) \geq f(j \mid \mathcal{Y}). \tag{1}$$

Submodularity ensures that a greedy algorithm achieves a bounded approximation factor when maximizing such a function [16].

To measure the similarity between a selected subset $\mathcal{S}$ and a target (or query) set $\mathcal{T}$, we use the Submodular Mutual Information (SMI) [17], [18]:

$$I_f(\mathcal{S}; \mathcal{T}) = f(\mathcal{S}) + f(\mathcal{T}) - f(\mathcal{S} \cup \mathcal{T}). \tag{2}$$

In our work, we use the Facility Location Mutual Information (FLMI) function [5]:

$$I_f(\mathcal{S}; \mathcal{T}) = \sum_{i \in \mathcal{T}} \max_{j \in \mathcal{S}} s_{ij} + \sum_{i \in \mathcal{S}} \max_{j \in \mathcal{T}} s_{ij}, \tag{3}$$

where s_{ij} denotes the cosine similarity between the MFCC feature vectors of utterances i and j. FLMI jointly captures both *query-relevance* and *representativeness* [5]. In our transductive, label-free selection setting, the unlabeled target audio serves as the query set.

We follow DITTO [10] by extracting utterance-level 39-d mean MFCC features, precomputing all candidate-query cosine similarities, and maximizing the FLMI function $I_f(\mathcal{S}; \mathcal{T})$ using a LazyGreedy and out-of-core exact greedy algorithm combination.

Let $\mathcal{C}$ denote the source candidate pool and $\mathcal{T}$ the unlabeled target query set. Running FLMI with $\mathcal{T}$ as the query yields a target-aware preselected subset $\mathcal{P} \subseteq \mathcal{C}$. Unless stated otherwise, we set the FLMI preselection budget to $|\mathcal{P}| = 30$k utterances. All subsequent selection, perturbation-based scoring, and fine-tuning operate on $\mathcal{P}$ or its subsets. We refer to the collection of all baseline and perturbed hypotheses constructed from $\mathcal{P}$ as the perturbation pool $\mathcal{Q}_{\mathcal{P}}$.

B. Perturbation-based Inference

For each utterance in the preselected subset $\mathcal{P}$, we decode a baseline hypothesis H_0 and a set of perturbed hypotheses $\{H_k\}_{k=1}^{K}$ generated by (i) model-parameter perturbation and (ii) input perturbation, forming the perturbation pool $\mathcal{Q}_{\mathcal{P}}$. Model perturbation injects zero-mean Gaussian noise into each parameter tensor W with scale $\alpha\,\mathrm{std}(W)$, where $\alpha \in \{0.01, 0.02, 0.03\}$; the baseline corresponds to $\alpha = 0.00$ (no parameter noise). Input perturbation (16 kHz) uses six variants: pitch-only shifts $\{-2, +1, +2\}$ semitones; and stretch–pitch pairs (atempo, pitch) $\in \{(0.90, -1), (0.95, -2), (0.95, -1)\}$. Overall, we generate 28 hypotheses per utterance before deduplication: 1 baseline (α=0.00 with no input perturbation), 6 input-perturbed hypotheses at α=0.00, and $3 \times 7 = 21$ model-perturbed hypotheses over $\alpha \in \{0.01, 0.02, 0.03\}$ applied to the original plus six input variants. We then deduplicate identical transcriptions, so K denotes the number of unique perturbed hypotheses kept (with $K \leq 27$). All hypotheses in $\mathcal{Q}_{\mathcal{P}}$ are subsequently scored by the multimodal, reference-free consistency metrics described in Sec. III-C and filtered by the acceptance rule in Sec. III-D to yield the final selected training subset.

Choice of perturbation types and scales. We select them on Common Voice with Whisper-base [19] (Sec. IV-E) and reuse the same configuration for other decoders/datasets without re-sweeping; we do not claim optimality on every backbone, but runtime selection still compares each hypothesis to its own baseline (Sec. III-D).

C. Sequence-level Embedding and WER Prediction

To obtain aligned speech–text representations, we use frozen pre-trained SONAR [6] encoders, which map audio and text inputs into a shared 1024-d language-agnostic embedding space.

We then train a lightweight reference-free WER predictor by concatenating the speech and text embeddings (2048-d input) and processing them through a 3-layer MLP (2048→600→32→1) with BatchNorm–Linear–ReLU–Dropout blocks. A sigmoid output with learnable temperature β constrains predictions to [0.01, 0.99].

Given the perturbation pool $\mathcal{Q}_{\mathcal{P}}$ for each utterance, we compute SONAR embeddings for the speech and for every hypothesis $H \in \mathcal{Q}_{\mathcal{P}}$. For each perturbed hypothesis H_k we obtain a bounded WER estimate $\hat{w}_k = f_\theta\big([\mathbf{h}^{\text{sp}} \| \mathbf{h}^{\text{txt}}(H_k)]\big)$. In parallel, we compute two alignment scores between the speech embedding $\mathbf{h}^{\text{sp}}$ and the text embedding $\mathbf{h}^{\text{txt}}(H_k)$: cosine similarity c_k and Euclidean distance d_k. We also compute the same alignment scores for the baseline H_0, yielding the baseline values used in Sec. III-D. The resulting per-hypothesis quality vectors $\{\hat{w}_k, c_k, d_k\}$ annotate $\mathcal{Q}_{\mathcal{P}}$ for selection:

We compute cosine similarity as:

$$c_k = \frac{\langle \mathbf{h}^{\text{sp}}, \mathbf{h}^{\text{txt}}(H_k) \rangle}{\|\mathbf{h}^{\text{sp}}\| \, \|\mathbf{h}^{\text{txt}}(H_k)\|}, \tag{4}$$

and Euclidean distance as:

$$d_k = \|\mathbf{h}^{\text{sp}} - \mathbf{h}^{\text{txt}}(H_k)\|_2. \tag{5}$$

These three quantities $\{\hat{w}, c, d\}$ constitute a quality vector that is passed to the selection procedure detailed in Sec. III-D.

D. Selection Strategy

For each utterance $u \in \mathcal{P}$, we compare its perturbed hypotheses $\{H_k\}$ against the baseline H_0. To do so, we define per-metric improvements. Let v_k and v_0 denote the metric values for H_k and H_0, respectively, where $v \in \{\hat{w}, c, d\}$. The improvement for metric v is:

$$\Delta_k^{(v)} = \begin{cases} v_0 - v_k, & v \in \{\hat{w}, d\} \\ v_k - v_0, & v \in \{c\} \end{cases}. \tag{6}$$

By aggregating only positive improvements over the perturbation pool $\mathcal{Q}_{\mathcal{P}}$ induced by the same FLMI pre-selected subset $\mathcal{P}$, we estimate empirical distributions and compute

percentile thresholds τ_v^p. The thresholds are calculated per metric, ensuring that the selection is adaptive to the score distribution of the current candidate pool.

At runtime, a hypothesis is accepted if its quality metrics exceed empirical percentile thresholds. We experiment with multiple selection rules: (i) *single-metric rules* using only one of $\{\Delta_k^{(\hat{w})}, \Delta_k^{(c)}, \Delta_k^{(d)}\}$, and (ii) a *conjunction rule* requiring both predicted-WER improvement and at least one alignment improvement:

$$\Delta_k^{(\hat{w})} \geq \tau_{\hat{w}}^p \ \wedge \ \left(\Delta_k^{(c)} \geq \tau_c^p \ \vee \ \Delta_k^{(d)} \geq \tau_d^p\right). \tag{7}$$

If multiple perturbations satisfy the rule for a given utterance, we keep the candidate with the largest predicted-WER improvement, i.e., the largest $\Delta_k^{(\hat{w})}$.

Applying the acceptance rule within $\mathcal{Q}_{\mathcal{P}}(u)$ for every $u \in \mathcal{P}$ yields at most one accepted hypothesis per utterance. The accepted set forms the final selected training subset $\mathcal{S} \subseteq \mathcal{P}$ with transcripts given by the retained hypotheses; utterances with no accepted hypothesis are excluded.

IV. Experiments

Organization. We first report fixed-epoch CRDNN/W2V2 results (Table IV) under Sec. IV-A–IV-E, including WER-predictor diagnostics (Table III). We then present variable-budget Paraformer sweeps in Sec. IV-F.

A. Experimental Setup

We adopt a matched-budget comparison protocol: all methods are trained under the same fine-tuning budget to isolate the effect of selection and pseudo-label quality. The CRDNN/Wav2Vec 2.0 (W2V2) experiments (Table IV) use a fixed 2-epoch budget; the Paraformer experiments (Sec. IV-F) provide a full budget sweep across hours.

Compared methods. We group baselines by their filtering signal:

Training references: Pretrained (no fine-tuning); *All-Ref* (supervised oracle with ground-truth labels); *All-Pseudo* (unfiltered pseudo-labels, null hypothesis).

PPL-based selection: PPL-P90 (text-only perplexity at matched percentile p=90); *PPL-SizeMatched* (same subset size as our method).

Multimodal selection (ours): Conf-P (conjunction rule); *Pred/Cos/Euc-Only* (single-metric variants); *Stable-Base-P* (baseline quality selection, Sec. IV-F).

Non-PPL baselines (Sec. IV-F): CER-Consistency (model-agreement across 3 ASRs); *SSL-WER-Classifier* (SVM on SSL features); *Random* (random sampling at matched hours).

This grouping guides our analysis: Sec. IV-E compares multimodal signals against PPL-based filtering under a fixed 2-epoch budget; Sec. IV-F provides a budget-sweep analysis including comparisons to both PPL and non-PPL baselines.

Perturbation selection. We select perturbation types and scales via an exploratory sweep on Common Voice using Whisper-base. Table I summarizes the explored and selected parameters for each perturbation type. For each perturbation configuration, we decode both the baseline hypothesis (no perturbation) and the perturbed hypothesis, compute WER against references, and report (i) the *improvement rate*, i.e., the fraction of utterances whose WER decreases relative to the baseline, and (ii) the average WER reduction over the improved utterances. We retain configurations with improvement rate $\geq$20% and select the final six input variants by ranking configurations by improvement rate. Specifically, the selected input variants are the top six by improvement rate among the retained configurations: three pitch-only shifts $\{-2, +1, +2\}$ semitones, and three stretch–pitch pairs $\{(0.90, -1), (0.95, -2), (0.95, -1)\}$.

TABLE I
Perturbation search space and selected parameters.

Type	Explored	Selected
Model noise α	{0.01, 0.02, 0.03, 0.04, 0.05}	{0.01, 0.02, 0.03}
Pitch shift (st)	{−2, −1, +1, +2}	{−2, +1, +2}
Time stretch	{0.90, 0.95, 1.05, 1.10}	{0.90, 0.95}
Additive noise	{0.005, 0.01, 0.02, 0.03}	— (excluded)

B. Dataset

We use three English corpora. **IndicTTS English subset** [20][1]. Originally designed for multilingual text-to-speech (TTS), the IndicTTS corpus includes professionally recorded and transcribed English read speech. Its English subset covers 13 Indian accents. We split this subset 80%/20% into a candidate pool $\mathcal{C}$ and a held-out target split; we sample 6,000 labeled dev utterances from $\mathcal{C}$ for early stopping and treat the 20% target audio as unlabeled for transductive querying. **CSTR VCTK Corpus** [21]. Read speech from 110 English speakers with diverse accents; used as an additional candidate pool. **L2-ARCTIC corpus** [22]. Non-native English with six L1 backgrounds; audio is resampled to 16 kHz (about 27 h total). For the L2-ARCTIC column in Table IV, $\mathcal{C}$= VCTK (cross-domain candidates). We randomly sample 6,000 L2-ARCTIC utterances for dev (early stopping) and use all remaining L2-ARCTIC audio as the unlabeled test set for queries and WER; dev/test are not speaker-disjoint. Paraformer uses a different L2-ARCTIC split and pool (Sec. IV-F).

We adopt a transductive, label-free protocol: unlabeled query audio drives FLMI preselection from $\mathcal{C}$; acceptance thresholds τ^p are computed *only* from scores on $\mathcal{Q}_{\mathcal{P}}$ (Sec. III-D), without using test transcripts. Training uses pseudo-labels; dev labels only support early stopping. All-Ref uses ground-truth only on the oracle fine-tuning subset; other labels are for WER evaluation only.

C. Models and Fine-tuning

We fine-tune three pretrained ASR backbones. We use SpeechBrain [23] for CRDNN and W2V2 fine-tuning. Table II summarizes training configurations.

(i) **CRDNN-RNNLM** (SpeechBrain/LibriSpeech [24]): 2 conv + 4 BiLSTM + 2 FC encoder with GRU decoder over 1k BPE. Fine-tuned with hybrid CTC/sequence loss (label smoothing),

[1]https://hf-mirror.com/SPRINGLab

TABLE II
TRAINING CONFIGURATIONS FOR FINE-TUNING.

Model	Epochs	LR	Batch	Optimizer
CRDNN [23]	2	3e−6	16	AdamW
W2V2 [23]	2	3e−6	16	Adam
Paraformer [27]	6	3e−5	64	Adam

TABLE III
WER-PREDICTOR PERFORMANCE ON PRESELECTED PERTURBED HYPOTHESIS SETS.

Dataset	Model	Pearson	Spearman	MAE	RMSE
IndicTTS [20]	CRDNN	0.86	0.77	0.09	0.13
	W2V2	0.84	0.84	0.12	0.17
VCTK [21]	CRDNN	0.85	0.85	0.14	0.21
	W2V2	0.94	0.81	0.09	0.13

AdamW + New-Bob scheduling; inference uses beam search with auxiliary RNNLM.

(ii) **Wav2Vec 2.0 + DNN** (SpeechBrain/Common Voice [25]): frozen wav2vec2 2.0 [26] frontend, 2 FC + GRU with dual CTC/attention heads. Fine-tuned with Adam + New-Bob; beam search at inference (no external LM).

(iii) **Paraformer** [27] (FunASR [28]): non-autoregressive model with Conformer encoder and bidirectional Transformer decoder. Fine-tuned from `paraformer-en` checkpoint with Adam + warmup scheduling.

D. Metric Computation Results

The WER predictor was trained on Common Voice [25] utterances transcribed by the Whisper-base [19] model using uncertainty inference, with reference WER values clamped to the range [0, 1]. We used the AdamW optimizer with cosine annealing, MSE loss, Xavier initialization, and a dropout rate of 0.3. We employed early stopping, halting training at 70 epochs. Although trained out of domain, it shows high correlation and low error on our target datasets (Table III), supporting its use as the core metric alongside cross-modal alignment for consistency filtering.

E. Fixed-epoch Evaluation

Experiments cover two models and two datasets (Table IV). **Reference rows (Pretrained, All-Pseudo, All-Ref).** Compared methods are defined in Sec. IV-A; here we interpret only these reference rows. All-Pseudo degrades substantially on both datasets compared to Pretrained, especially for CRDNN, indicating that unfiltered pseudo-labels introduce harmful noise. All-Ref shows that in-domain labeled data (IndicTTS) provides positive gains, while out-of-domain labeled data (L2-ARCTIC with VCTK source) even hurts performance compared to zero-shot, suggesting negative transfer under domain mismatch.

Key observations. (i) Effectiveness and model dependence (IndicTTS). Consistency-filtered subsets (Conf-P^p) yield substantial gains over unfiltered pseudo-label training, especially for the weaker CRDNN backbone, which is more susceptible to pseudo-label noise. With W2V2, using 1.5k selected utterances from a 30k pool attains 10.91% WER, close to 10.45% from 30k supervised labels. Across the pre-specified percentile sweep, WERs vary only slightly (10.91–11.27%), suggesting perturbation-consistent utterances provide high-quality training signals in this in-domain setting.

(ii) Behavior under domain shift (L2-ARCTIC). Under strong accent shift, consistency-filtered subsets avoid the degradation observed with unfiltered pseudo-labels: for CRDNN, Conf-P95 reaches 22.00% while Cos/Euc achieve 20.03%/21.93%; for W2V2, results cluster tightly around the zero-shot baseline (15.19–15.33%). Given the candidate domain (VCTK) is acoustically simpler than L2-ARCTIC, extracting helpful pseudo-labels is non-trivial, yet filtered subsets provide measurable benefit, highlighting stability rather than degradation.

(iii) Multimodal signals vs. text-only filtering. Multimodal variants generally outperform text-only PPL filtering for CRDNN and remain competitive for W2V2 (Table IV, PPL Contrast rows), supporting that speech–text alignment and predicted WER capture quality information complementary to lexical perplexity.

Note on ablations. Fixing the same percentile threshold p can yield different selected subset sizes across rules (Table IV); we therefore interpret single-metric comparisons primarily as diagnostics of signal informativeness. Among single-metric variants, alignment-based signals (Cos-Only, Euc-Only) often perform competitively with the conjunction rule, suggesting that the primary value lies in the multimodal signals rather than the specific combination logic. This observation also motivates our use of matched audio-hours (rather than matched utterance counts) in Sec. IV-F for more equitable budget comparisons.

In summary, under a fixed 2-epoch fine-tuning budget, selection benefits are modulated by model capacity and domain difficulty: larger absolute gains arise in-domain (especially for CRDNN), while under strong shift the effect is primarily to mitigate error amplification without harming zero-shot performance. Across all settings, multimodal signals generally outperform text-only PPL, validating their utility for pseudo-label quality estimation in accent adaptation.

F. Variable-budget Evaluation

To provide a complementary, budget-sweep validation on a stronger ASR backbone, we report additional experiments using Paraformer on English test sets (L2-ARCTIC and IndicTTS English). These experiments isolate the effect of the post-selection module by applying it directly to a fixed candidate pool $\mathcal{C}$ without the FLMI pre-selection stage and analyzing WER under matched-hour budgets. We present three complementary views: (i) a WER–hours sweep to characterize budget-efficiency and compare multimodal filtering against text-only PPL, (ii) a random-sampling sanity check at matched hours, and (iii) a comparison to recent non-PPL selection baselines under matched-hour budgets. All-Ref serves as a supervised oracle, and All-Pseudo uses unfiltered pseudo-transcriptions.

TABLE IV

WER (%) UNDER A MATCHED FINE-TUNING BUDGET (2 EPOCHS) FOR CRDNN AND W2V2 IN AN IN-DOMAIN AND A CROSS-DOMAIN SETTING. SELECTION IS TRANSDUCTIVE AND LABEL-FREE; ALL-REF IS A SUPERVISED ORACLE. UNLESS NOTED, WE USE p=90 (SEE SEC. III-D AND SEC. IV-E). THE SINGLE-METRIC BLOCK OMITS CONF-P90 BECAUSE IT MATCHES THE CONJUNCTION CONF-P90 ROW ABOVE; THE PPL CONTRAST BLOCK LIKEWISE OMITS IT AND LISTS PPL BASELINES ONLY.

Stage / Strategy	Method	IndicTTS				L2-ARCTIC			
		CRDNN	Utts	W2V2	Utts	CRDNN	Utts	W2V2	Utts
Pre-Train	Pretrained (no FT)	14.90	-	12.03	-	22.97	-	15.31	-
Baseline	All-Ref	9.10	30k	10.45	30k	25.96	30k	16.42	30k
	All-Pseudo	30.11	30k	12.52	30k	32.73	30k	17.24	30k
Consistency	Conf-P95	12.72	911	11.23	282	**22.00**	687	15.32	331
	Conf-P90	**12.70**	1751	11.27	712	22.22	1225	15.32	485
	Conf-P80	13.08	3504	**10.91**	1519	22.13	2318	**15.19**	882
	Conf-P70	15.13	5190	11.11	2370	22.50	3505	15.21	1268
Single-Metric	Pred-Only	13.25	2317	11.30	1115	22.05	1525	**15.19**	571
	Cos-Only	12.87	2876	**10.90**	1014	**20.03**	1898	15.33	638
	Euc-Only	12.83	3165	**10.90**	1088	21.93	1949	15.26	614
PPL Contrast	PPL-P90	15.44	1499	11.44	782	25.18	1269	**15.31**	494
	PPL-SizeMatched	15.25	1751	11.29	712	24.71	1225	15.32	485

Candidate pool and L2 splits (vs. Table IV). IndicTTS uses the same 30k FLMI-fixed $\mathcal{C}$ without re-running FLMI. On L2-ARCTIC, Paraformer selects, for each L1 accent, one speaker for dev (validation) and one other speaker for test; $\mathcal{C}$ pools utterances from all remaining speakers (in-domain). Sec. IV-E uses $\mathcal{C}=$ VCTK with random 6k/rest L2 dev/test without a speaker-disjoint constraint. The two L2 setups differ: in-domain Paraformer vs. cross-domain Table IV.

Purification-oriented variants (additional strategies). Besides Conf-P^p (which selects utterances where perturbation yields improved hypotheses), we include two variants to disentangle whether gains come from perturbation-improvement signals or simply from selecting high-quality baseline pseudo-labels—similar to the filtering approach in prior work [11].

(i) *Stable-Base-P^p*: selects utterances whose *baseline* hypothesis H_0 (without perturbation) is already high-quality, analogous to prior pseudo-label filtering methods that retain only high-confidence transcriptions. Unlike Conf-P, this strategy does not require improvement over the baseline; instead, it directly thresholds the baseline quality scores. Formally, an utterance u is selected if:

$$\hat{w}_0 \le \tau^p_{\hat{w}} \ \wedge \ \big(c_0 \ge \tau^{100-p}_c \ \vee \ d_0 \le \tau^p_d\big), \tag{8}$$

where $\hat{w}_0$, c_0, and d_0 are the predicted WER, cosine similarity, and Euclidean distance for the baseline hypothesis, and τ^p denotes the p-th percentile threshold computed from the *baseline hypothesis distribution* over $\mathcal{P}$ (not the improvement distribution). For lower-is-better metrics ($\hat{w}$, d), we use the p-th percentile; for higher-is-better (c), we use the $(100-p)$-th percentile.

(ii) *Conf-Stable-p_1-p_2*: inspired by curriculum learning, merges a Conf-P^{p_1} subset (perturbation-improved hypotheses) with a Stable-Base-P^{p_2} subset (baseline high-quality hypotheses):

$$\mathcal{S}^{p_1,p_2}_{\text{conf-stable}} = \mathrm{Merge}\big(\mathcal{S}^{p_1}_{\text{conf}}, \mathcal{S}^{p_2}_{\text{stable}}\big). \tag{9}$$

For example, Conf-Stable-P70-P50 combines Conf-P70 and Stable-Base-P50. The operator $\mathrm{Merge}(\cdot)$ concatenates the two selected subsets; if an utterance appears in both, it is included twice, effectively upweighting it. This variant tests whether combining both selection paradigms—perturbation-improved and baseline-quality—yields additional benefit.

Results show that both strategies can be effective, suggesting that the multimodal signals themselves—not the perturbation mechanism alone—are the key drivers of quality estimation.

Budget-sweep analysis (WER–hours trade-off). Fig. 2 summarizes full percentile sweeps as WER versus selected hours (log-scale). We extend the percentile range to $p \in \{50, 60, 70, 80, 90, 95\}$ for finer-grained budget analysis. Three key observations emerge from the sweep:

(i) **Multimodal signals generally outperform text-only PPL**: Across both datasets, the PPL curve tends to lie above multimodal selection curves at comparable hour budgets. For example, at the p=60 operating point on L2-ARCTIC, Conf-P60 achieves 17.41% WER at 1.867 h, compared to PPL-P60's 17.59% at 1.851 h; on IndicTTS, Conf-P60 reaches 7.11% at 4.884 h versus PPL-P60's 7.19% at 4.766 h. This suggests that speech–text alignment signals capture quality information complementary to lexical perplexity.

(ii) **Data efficiency**: On L2-ARCTIC, the merged variant Conf-Stable-P70-P50 (Eq. 9) achieves 16.86% WER at 7.272 h (51.9% of All-Pseudo's 14.017 h), outperforming All-Pseudo's 17.02%, suggesting data purification is possible in this setting. On IndicTTS, where All-Pseudo is already strong (6.90% WER), post-selection subsets offer a trade-off between data budget and accuracy rather than strict WER improvement—e.g., Euc-Only-P50 reaches 7.08% with only 15.5% of the hours.

(iii) **Relative stability across signals**: Among multimodal variants, Euc-Only tends to perform competitively across hour budgets; the differences among Conf-P, Pred-Only, Cos-Only, and Euc-Only are generally small, indicating that all proposed multimodal signals tend to outperform text-only PPL and practitioners can choose based on computational convenience.

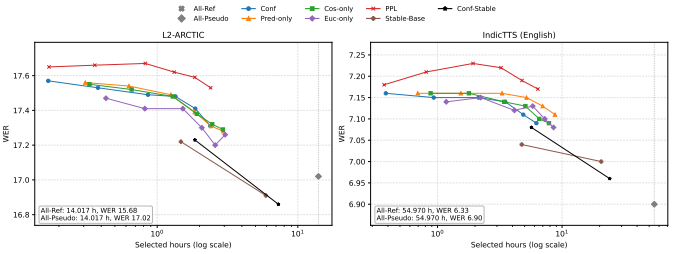

Fig. 2. Paraformer post-selection: WER versus selected hours (log scale).

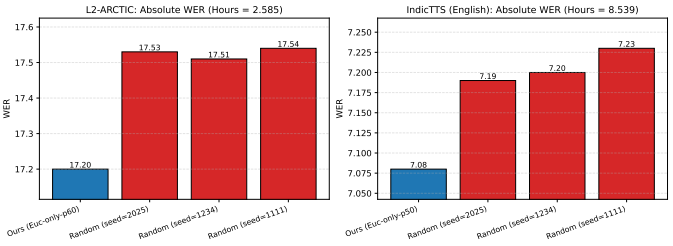

Fig. 3. Performance comparison under matched audio-hour budgets: multimodal post-selection vs. random subsets (multiple seeds).

Because the mapping from percentile p to selected hours depends on utterance-length and score distributions, we interpret these trends as directional observations rather than claiming a single globally optimal p.

Sanity check (random baseline at matched hours). Multimodal post-selection generally outperforms random sampling at matched durations, demonstrating its ability to avoid low-quality pseudo-labels. Fig. 3 compares our method to multiple random controls (different random seeds) under representative budgets of roughly 15–20% of the full pseudo-labeled hours (All-Pseudo). We select the operating point whose hours most closely match the target budget from our percentile sweep: Euc-Only-P60 on L2-ARCTIC (17.20 WER at 2.585 h) and Euc-Only-P50 on IndicTTS (7.08 WER at 8.539 h). On L2-ARCTIC, the selected subset outperforms three random controls at the same duration (17.51–17.54 WER). On IndicTTS, the selected subset similarly improves over random subsets matched by hours (7.19–7.23 WER). These comparisons reduce the chance that improvements arise from accidental subsampling.

Comparison to recent non-PPL selection baselines. Our quality-based selection is competitive with model-agreement or SSL-classification baselines under matched-hour budgets. We compare to two baselines from [11]: (i) CER-Consistency, which selects utterances based on the average CER consistency across three ASR hypotheses (Paraformer, Zipformer, and Parakeet), keeping segments with $\mathrm{CER}_{\mathrm{avg}}(s) < \tau$ (e.g., 5%) where $\mathrm{CER}_{\mathrm{avg}}(s) = \frac{1}{3}\left(\mathrm{CER}_{\mathrm{p-z}}(s) + \mathrm{CER}_{\mathrm{p-k}}(s) + \mathrm{CER}_{\mathrm{z-k}}(s)\right)$, and (ii) SSL-WER-Classifier, which uses an SVM on SSL features to predict WER and filter low-quality pseudo-labels (below 0.5 as high-quality; above 0.5 as low-quality). Since these baselines filter based on baseline hypothesis quality rather than perturbation-improvement, we use Stable-Base-P50 as a representative of our approach for a fair comparison. Under the same hour budget on L2-ARCTIC (Table V), Stable-Base-P50 outperforms both baselines across multiple random seeds, suggesting that multimodal quality signals can provide competitive or better selection than model-agreement or SSL-feature classification.

TABLE V
COMPARISON TO RECENT SELECTION BASELINES UNDER MATCHED HOURS (L2-ARCTIC, PARAFORMER).

Method	Hour	WER
Stable-Base-P50 (ours)	5.923	**16.91**
CER-Consistency (seed 2025)	5.923	16.99
CER-Consistency (seed 1234)	5.923	17.06
CER-Consistency (seed 1111)	5.923	17.02
SSL-WER-Classifier (seed 2025)	5.923	17.22
SSL-WER-Classifier (seed 1234)	5.923	17.21
SSL-WER-Classifier (seed 1111)	5.923	17.23

V. CONCLUSIONS

Our experiments show that multimodal consistency signals—speech–text alignment in SONAR space and predicted WER—provide a practical reference-free criterion to filter pseudo-labels for ASR accent adaptation. Our main finding is that these multimodal signals provide more reliable quality estimation than text-only perplexity across different backbones and domain conditions. Absolute WER gains can be modest on stronger backbones; the practical benefit is often to suppress harmful pseudo-label noise and maintain stability under accent or domain mismatch rather than uniformly large improvements. In an in-domain setting, $\sim$1.5k selected utterances achieve 10.91% WER, approaching the 10.45% from 30k supervised labels. In a cross-domain setting with a mismatched candidate pool, consistency filtering avoids degradation that unfiltered pseudo-labels cause. Matched-hour experiments on a stronger ASR backbone confirm that the post-selection module generalizes across backbones and outperforms random sampling and recent selection baselines under matched budgets. *Limitations.* p, frozen SONAR, and the fixed perturbation pool are pragmatic defaults: we do not retune p per corpus or exhaustively ablate alternative encoders or perturbation sets. Beyond English read-speech corpora, multilingual, noisy-channel, and long-tail conditions are not evaluated. Future work includes adaptive thresholds, continual adaptation, and inductive extensions such as recalibrating acceptance when deployment conditions drift from the query audio or when selection is decoupled from later test traffic.

Source code is available at https://github.com/hakuna29/mcds-asr.

ACKNOWLEDGMENT

This work was supported in part by the Xinjiang Uygur Autonomous Region Key R&D Task Program under Grant No. 2025B03042-1, and in part by the Key Project of the National Language Commission's 14th Five-Year Scientific Research Plan under Grant No. ZDI145-172.

REFERENCES

[1] B. R. Bartoldson, B. Kailkhura, and D. W. Blalock, "Compute-efficient deep learning: Algorithmic trends and opportunities," *ArXiv*, vol. abs/2210.06640, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252873115

[2] J. Shor, D. Emanuel, O. Lang, O. Tuval, M. Brenner, J. Cattiau, F. Vieira, M. McNally, T. Charbonneau, M. Nollstadt *et al.*, "Personalizing asr for dysarthric and accented speech with limited data," *arXiv preprint arXiv:1907.13511*, 2019.

[3] K. C. Sim, P. Zadrazil, and F. Beaufays, "An investigation into on-device personalization of end-to-end automatic speech recognition models," *arXiv preprint arXiv:1909.06678*, 2019.

[4] A. Awasthi, A. Kansal, S. Sarawagi, and P. Jyothi, "Error-driven fixed-budget asr personalization for accented speakers," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7033–7037.

[5] S. Kothawade, V. Kaushal, G. Ramakrishnan, J. Bilmes, and R. Iyer, "Prism: A rich class of parameterized submodular information measures for guided data subset selection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 9, 2022, pp. 10 238–10 246.

[6] P.-A. Duquenne, H. Schwenk, and B. Sagot, "Sonar: sentence-level multimodal and language-agnostic representations," *arXiv preprint arXiv:2308.11466*, 2023.

[7] N. Lagos and I. Calapodescu, "Unsupervised multi-domain data selection for asr fine-tuning," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 711–10 715.

[8] Z. Zheng, Z. Ma, Y. Wang, and X. Chen, "Unsupervised active learning: Optimizing labeling cost-effectiveness for automatic speech recognition," *arXiv preprint arXiv:2308.14814*, 2023.

[9] P. Rangappa, J. Zuluaga-Gomez, S. Madikeri, A. Carofilis, J. Prakash, S. Burdisso, S. Kumar, E. Villatoro-Tello, I. Nigmatulina, P. Motlicek *et al.*, "Speech data selection for efficient asr fine-tuning using domain classifier and pseudo-label filtering," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[10] S. Kothawade, A. Mekala, M. Kothyari, R. Iyer, G. Ramakrishnan, P. Jyothi *et al.*, "Ditto: Data-efficient and fair targeted subset selection for asr accent adaptation," *arXiv preprint arXiv:2110.04908*, 2021.

[11] P. Rangappa, A. Carofilis, J. Prakash, S. Kumar, S. Burdisso, S. Madikeri, E. Villatoro-Tello, B. Sharma, P. Motlicek, K. Hacioglu *et al.*, "Efficient data selection for domain adaptation of asr using pseudo-labels and multi-stage filtering," *arXiv preprint arXiv:2506.03681*, 2025.

[12] K. A. Yuksel, T. Ferreira, A. Gunduz, M. Al-Badrashiny, and G. Javadi, "A reference-less quality metric for automatic speech recognition via contrastive-learning of a multi-language model with self-supervision," in *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2023, pp. 1–5.

[13] C. Park, C. Lu, M. Chen, and T. Hain, "Fast word error rate estimation using self-supervised representations for speech and text," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[14] C. Park, M. Chen, and T. Hain, "Automatic speech recognition system-independent word error rate estimation," *arXiv preprint arXiv:2404.16743*, 2024.

[15] S. Fujishige, *Submodular functions and optimization*. Elsevier, 2005, vol. 58.

[16] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—i," *Mathematical programming*, vol. 14, no. 1, pp. 265–294, 1978.

[17] A. Gupta and R. Levin, "The online submodular cover problem," in *Proceedings of the Thirty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '20. USA: Society for Industrial and Applied Mathematics, 2020, p. 1525–1537.

[18] R. Iyer and J. Bilmes, "A memoization framework for scaling submodular optimization to large scale problems," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 2340–2349.

[19] 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.

[20] S. T. Consortium, Hema A Murthy, and S Umesh, "Indic TTS: A text-to-speech database for indian languages," 2023. [Online]. Available: https://www.iitm.ac.in/donlab/indictts/

[21] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)," [Sound]. University of Edinburgh, The Centre for Speech Technology Research (CSTR), 2019. [Online]. Available: https://doi.org/10.7488/ds/2645

[22] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-arctic: A non-native english speech corpus," in *Proc. Interspeech*, 2018, p. 2783–2787. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1110

[23] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong *et al.*, "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[24] 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.

[25] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the twelfth language resources and evaluation conference*, 2020, pp. 4218–4222.

[26] 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.

[27] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, "Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition," in *Proc. Interspeech*, 2022, pp. 2063–2067.

[28] Z. Gao, Z. Li, J. Wang, H. Luo, X. Shi, M. Chen, Y. Li, L. Zuo, Z. Du, Z. Xiao *et al.*, "Funasr: A fundamental end-to-end speech recognition toolkit," *arXiv preprint arXiv:2305.11013*, 2023.