CUMULATIVE ACOUSTIC VARIATION FOR TRAINING-FREE ASR DATA PRUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Training-data pruning for automatic speech recognition (ASR) aims to reduce training-data requirements, yet model-dependent selection criteria can incur substantial computational overhead. To address this, we propose a model-free acoustic pruning approach that ranks training utterances using cumulative statistics computed directly from speech signals. Specifically, we develop three utterance-level criteria based on accumulated spectral entropy, spectral flux, and log-Mel temporal variation, requiring neither pruning-specific model training nor learned-model inference. Under a fixed utterance-count budget, these criteria retain both the extent and the local acoustic characteristics of each utterance through cumulative scoring. Experiments with Conformer on LibriSpeech show that the proposed acoustic criteria achieve the best performance at 5% and 1% retention and remain competitive with the best-performing Teacher–Student criterion at 30% and 10% retention, while reducing data-selection time from 13.73 hours to less than 0.5 hours without requiring GPUs. In particular, Log-Mel Temporal Variation achieves the lowest WER at both 5% and 1% retention; at 5%, it reduces WER relative to random selection by 26.1% and 15.3% on test-clean and test-other, respectively. Ablations further support the use of cumulative rather than mean aggregation under an utterance-count budget. These results demonstrate that simple cumulative acoustic statistics provide an efficient and effective alternative to model-dependent criteria for ASR training-data pruning, with especially strong performance under highly restrictive retention budgets.

1 INTRODUCTION

Training-data pruning offers a way to reduce the number of training examples while preserving as much downstream performance as possible (Toneva et al., 2018; Paul et al., 2021). For automatic speech recognition (ASR), this requires selecting a subset of paired speech utterances and transcripts (Mittal et al., 2022; Xiao et al., 2024). The central challenge is to identify which utterances should be retained under a limited data budget. An utterance can be valuable for different reasons: its informative linguistic content, its acoustic realization which exposes useful variation, or its difficulty that reveals limitations of the current model. These possibilities motivate the investigation of different selection signals and their effectiveness as the retained subset becomes smaller.

Selection quality, however, is only one aspect of a practical pruning method. Criteria based on early-training statistics (Paul et al., 2021), relative-difficulty comparisons (Mindermann et al., 2022), or pretrained representations (Aldeneh et al., 2026) require model training, inference, or full-dataset feature extraction before a subset can be constructed. Such computations introduce an additional cost that should be considered alongside downstream recognition performance (Coleman et al., 2019; Dai et al., 2025). In particular, access to a powerful scoring model does not by itself establish that the resulting subset will provide the best performance–efficiency trade-off. This motivates a question: can useful ASR training utterances be identified directly from the speech signal, without training or running a learned scoring model?

Speech signals provide a natural starting point for this investigation. Their spectral distributions and temporal evolution can be characterized through deterministic acoustic-feature computations (Misra et al., 2004; Nam et al., 2015). These statistics describe properties of the observed signal that may provide useful selection cues alongside linguistic or model-derived information. As illustrated

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

in Figure 1, utterances of similar duration exhibit a broad range of mean frame-to-frame log-Mel variation. This variability motivates investigating local acoustic variation as a selection signal beyond utterance length alone. At the same time, variable utterance lengths make score aggregation an important design choice. In our setting, the pruning budget constrains the number of retained utterances rather than their total duration. Each utterance therefore occupies one selection slot, although the amount of audio it contains can differ substantially. A mean statistic describes an average local property, whereas a cumulative statistic retains both that property and the extent over which it is observed. This motivates the hypothesis that cumulative acoustic statistics can serve as effective utterance-level selection criteria under an utterance-count budget.

To address the problems mentioned above, we propose a model-free acoustic pruning approach built on three cumulative scoring criteria: Accumulated Spectral Entropy, Accumulated Spectral Flux, and Log-Mel Temporal Variation. Accumulated Spectral Entropy measures spectral dispersion within individual frames and sums it over the utterance. Accumulated Spectral Flux measures changes in spectral shape across adjacent frames, while Log-Mel Temporal Variation accumulates frame-to-frame changes in the log-Mel representation. Together, these criteria capture complementary aspects of spectral structure and temporal variation. Each maps an utterance directly to a scalar score. The scoring procedure requires neither pruning-specific model training nor inference from a pretrained learned model.

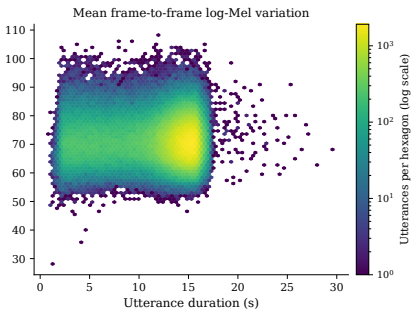


Figure 1: Joint distribution of utterance duration and mean frame-to-frame log-Mel variation in the LibriSpeech training set. Color indicates the number of utterances per hexagonal bin on a logarithmic scale.

Our experiments demonstrate that direct acoustic scoring can provide both recognition and computational benefits. Log-Mel Temporal Variation achieves the strongest recognition performance among the evaluated criteria under aggressive pruning, while the acoustic family remains competitive with the strongest model-dependent criterion at larger retention budgets. These results are obtained with substantially lower selection overhead and without GPU resources for scoring. The findings establish cumulative acoustic statistics as a practical basis for ASR training-data selection and show that more computationally demanding scoring mechanisms do not necessarily produce more effective training subsets.

Our main contributions are as follows:

- We propose a model-free acoustic pruning approach for ASR based on accumulated spectral entropy, spectral flux, and log-Mel temporal variation. These criteria rank training utterances directly from speech signals without requiring a learned scoring model.
- We systematically compare the performance and time cost of selection criteria derived from different information sources under a common LibriSpeech–Conformer setup.
- We provide empirical support for cumulative scoring under an utterance-count budget, showing its advantage over mean aggregation while demonstrating that selection performance still depends on the audio information captured by the scoring mechanisms.

2 Related Work

2.1 Data Selection and Pruning for ASR

Dataset pruning selects a subset of an existing training dataset with the aim of reducing data requirements while preserving downstream generalization performance(Toneva et al., 2018; Welling, 2009; Xia et al., 2022; Margatina et al., 2021; Tang et al., 2023; Okanovic et al., 2024; Xu et al., 2023; Dolatabadi et al., 2022; Killamsetty et al., 2021; Mirzasoleiman et al., 2020; Huang et al., 2023; Nguyen & He, 2025). For ASR, this involves identifying speech utterances that support effective recognizer training under a limited data budget. (Mittal et al., 2022) propose Partitioned Gradient Matching to make gradient-based subset selection feasible for RNN-T training on large speech

datasets. (Xiao et al., 2024) investigate dynamic data pruning and introduce speech-specific pruning granularities beyond complete utterances. Representation-based approaches provide another selection strategy. (Aldeneh et al., 2026) use embeddings describing speaker, phonetic, and semantic characteristics to balance target-domain relevance and subset diversity, combining multiple representation spaces through score-level fusion. These studies cover complementary settings, including training-time data selection and target-domain subset construction.

2.2 Model-Based Sample Scoring

Language-model scores provide an established basis for corpus selection. (Moore & Lewis, 2010) select language-model training text using the difference between in-domain and background-model cross-entropies. Their criterion normalizes scores by sentence length and prioritizes text relevant to the target domain.

Training behavior offers another source of sample-level information. (Paul et al., 2021) introduce GraNd and EL2N scores based on gradients and prediction errors early in training, demonstrating their usefulness for identifying examples to retain. RHO-LOSS (Mindermann et al., 2022) instead prioritizes examples through the difference between the current model loss and an estimate of irreducible loss. This relative-loss formulation distinguishes examples that remain useful to learn from those that are already learned or difficult to learn. In ASR active learning, (Luo et al., 2021) use predicted CTC and attention losses to select unlabeled utterances for annotation. Although these approaches differ in their objectives and selection protocols, they illustrate how linguistic predictability and model behavior can guide sample selection.

2.3 Model-Free Pruning and Acoustic Statistics

Task-specific information can support pruning without learned scoring models. For instance segmentation, (Dai et al., 2025) construct scores from the shape and class information in instance annotations, incorporating scale and class normalization. In speech and audio processing, spectral statistics also have established applications. (Misra et al., 2004) use spectral entropy as an additional feature for robust ASR. (Nam et al., 2015) employ a spectral-flux-derived onset function to guide local spectro-temporal patch sampling for music auto-tagging.

Sequence length provides a further selection cue. (Mittal et al., 2022) include a longest-utterance baseline in their ASR subset-selection experiments. In self-supervised speech pre-training, (Whetten et al., 2026) find that prioritizing long utterances can outperform the evaluated diversity-based strategies. These findings motivate attention to sequence extent when interpreting cumulative scores. Building on these research directions, our work develops cumulative acoustic statistics into utterance-level criteria for static ASR training-data pruning without learned scoring models.

3 Method

3.1 Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a complete ASR training set containing N utterances, where x_i is the speech signal of the i-th utterance and y_i is its corresponding ground-truth transcript. We consider utterance-level training-data pruning, where each utterance together with its transcript is treated as one indivisible selection unit.

Given a target retention ratio $r \in (0, 1]$, the goal is to construct a subset

$$\mathcal{D}_r \subseteq \mathcal{D}, \qquad |\mathcal{D}_r| = B, \qquad B = \lfloor rN \rfloor, \tag{1}$$

such that an ASR model trained on $\mathcal{D}_r$ retains as much downstream recognition performance as possible despite using substantially fewer training utterances. Formally, under a fixed training procedure, the ideal subset can be expressed as

$$\mathcal{D}_r^* = \arg \min_{\substack{\mathcal{S} \subseteq \mathcal{D} \\ |\mathcal{S}| = B}} \mathrm{WER}\left(f_{\theta(\mathcal{S})}\right), \tag{2}$$

where $f_{\theta(\mathcal{S})}$ denotes the ASR model with parameters $\theta(\mathcal{S})$ obtained by training on subset $\mathcal{S}$, and $\mathrm{WER}(\cdot)$ denotes the downstream word error rate.

Most criteria studied in this work assign each utterance a scalar importance score based on its transcript, speech signal, or model-derived statistics. Unless otherwise specified, larger scores indicate higher retention priority, and the B highest-ranked utterances are selected. We also consider a representation-based criterion that performs structured subset construction through clustering rather than a single global ranking.

3.2 Model-Based Criteria

LM Surprisal. Motivated by (Moore & Lewis, 2010), we measure transcript unpredictability using a pretrained autoregressive language model. For an utterance with M_i transcript tokens, we consider both mean and cumulative surprisal:

$$S_i^{\mathrm{MeanLM}} = -\frac{1}{M_i}\sum_{j=1}^{M_i}\log p_{\mathrm{LM}}(y_{i,j} \mid y_{i,<j}),$$
$$S_i^{\mathrm{CumLM}} = -\sum_{j=1}^{M_i}\log p_{\mathrm{LM}}(y_{i,j} \mid y_{i,<j}). \tag{3}$$

The former measures length-normalized linguistic unpredictability (*Surprisal*), while the latter retains transcript length to capture accumulated linguistic information (*Surprisal-Sum*). Higher scores receive higher retention priority.

CTC–Attention Loss Disagreement (Disagreement). Hybrid CTC/Attention ASR models optimize two sequence objectives that characterize an utterance from different modeling perspectives. To capture their disagreement during the early stage of training, we use an ASR model trained for only two epochs, whose parameters remain far from full convergence. For utterance i, let L_i^{CTC} and L_i^{Att} denote the sample-level CTC and attention losses produced by this early-stage model. Following Luo et al. (2021), We quantify their disagreement as

$$S_i^{\mathrm{Dis}} = \left|L_i^{\mathrm{CTC}} - L_i^{\mathrm{Att}}\right|. \tag{4}$$

The two branch losses are used directly, without applying the weighting coefficients used to combine them during ASR training. A large score indicates that, at an early stage of optimization, the CTC and attention objectives assign substantially different difficulty to the same utterance. Such utterances expose complementary modeling challenges across the two objectives.

Audio–Text Joint Embedding Selection (Joint). Motivated by (Aldeneh et al., 2026), we further evaluate a representation-based strategy that preserves coverage of the joint audio–text distribution. For each utterance, pretrained speech and text encoders produce utterance-level embeddings via masked mean pooling. Each modality is independently L2-normalized, reduced by modality-specific PCA, normalized again, and concatenated:

$$z_i = \left[\mathrm{norm}\Big(\mathrm{PCA}_a(e_i^{(a)})\Big)\,;\,\mathrm{norm}\Big(\mathrm{PCA}_t(e_i^{(t)})\Big)\right]. \tag{5}$$

We apply K-means with K-means++ initialization to the joint representations. For cluster c containing n_c of N utterances, a retention budget of B is allocated approximately in proportion to cluster size:

$$q_c \approx B\frac{n_c}{N}, \qquad \sum_c q_c = B. \tag{6}$$

Within each cluster, the q_c utterances closest to its centroid in Euclidean distance are retained. Thus, proportional allocation preserves coverage across the joint representation space, while centroid-priority selection favors representative samples within each region.

3.3 Training-Free Acoustic Pruning

The comparison criteria in Sec. 3.2 rely on transcripts, learned representations, or model-derived quantities. We additionally investigate a family of training-free acoustic pruning criteria that operate directly on the speech signal. These criteria require neither pruning-specific model training nor inference from a learned foundation model. Instead, each utterance is mapped to a scalar score using deterministic acoustic-feature statistics.

Accumulated Spectral Entropy (SpecEntropy) We first characterize the cumulative spectral dispersion of an utterance. Given the short-time Fourier transform (STFT) coefficient $X_{t,f}$ at frame t and frequency bin f, we compute the power spectrum

$$P_{t,f} = |X_{t,f}|^2. \tag{7}$$

For each frame, the power values are normalized across frequency to form a discrete spectral distribution:

$$p_{t,f} = \frac{P_{t,f}}{\sum_{f'} P_{t,f'} + \epsilon}, \tag{8}$$

where ϵ is a small constant for numerical stability.

The spectral entropy of frame t is then

$$H_t = -\sum_f p_{t,f} \log\left(p_{t,f} + \epsilon\right). \tag{9}$$

A large H_t indicates that the spectral energy is distributed broadly across frequency bins, whereas a small value indicates that the energy is concentrated in a smaller number of frequencies.

We define the utterance-level score by accumulating the frame-wise entropy over the complete utterance:

$$S_i^{\mathrm{SE}} = \sum_{t=1}^{T_i} H_t, \tag{10}$$

where T_i denotes the number of frames in utterance i. Utterances with larger accumulated spectral entropy are assigned higher retention priority.

Accumulated Spectral Flux (SpecFlux) Spectral entropy describes the distribution of energy within each frame but does not directly measure how the spectrum evolves through time. To capture temporal spectral dynamics, we define a cumulative spectral-flux criterion.

Let A_t denote the STFT magnitude spectrum at frame t. To reduce the influence of absolute signal magnitude, each frame is independently L2-normalized:

$$\hat{A}_t = \frac{A_t}{\|A_t\|_2 + \epsilon}. \tag{11}$$

We measure the spectral change between two adjacent frames using their Euclidean distance,

$$F_t = \left\| \hat{A}_t - \hat{A}_{t-1} \right\|_2. \tag{12}$$

The utterance-level score is the accumulated frame-to-frame variation:

$$S_i^{\mathrm{SF}} = \sum_{t=2}^{T_i} \left\| \hat{A}_t - \hat{A}_{t-1} \right\|_2. \tag{13}$$

This criterion emphasizes changes in spectral shape rather than absolute loudness. Speech transitions and other acoustic dynamics can produce substantial spectral evolution over time, thereby contributing to a higher selection score for the utterance.

Log-Mel Temporal Variation (LogMel-TV) We instantiate the same cumulative temporal-variation principle directly in an ASR-oriented acoustic feature space. Let $M_i \in \mathbb{R}^{T_i \times F}$ denote the log-Mel representation of utterance i, where $M_{i,t}$ is the F-dimensional feature vector at frame t.

For two adjacent frames, we measure their variation using the L1 distance:

$$D_{i,t} = \|M_{i,t} - M_{i,t-1}\|_1 = \sum_f |M_{i,t,f} - M_{i,t-1,f}|. \tag{14}$$

The Log-Mel Temporal Variation score is

$$S_i^{\mathrm{LMTV}} = \sum_{t=2}^{T_i} \|M_{i,t} - M_{i,t-1}\|_1. \tag{15}$$

Table 1: WER (%) on LibriSpeech test-clean and test-other under different retention ratios. Lower is better. Bold indicates the best pruning result at each retention ratio.

Method	test-clean					test-other				
	100%	**30%**	**10%**	**5%**	**1%**	**100%**	**30%**	**10%**	**5%**	**1%**
Full Data	2.4	–	–	–	–	5.5	–	–	–	–
Random	–	3.5	6.6	14.2	93.3	–	8.1	14.2	24.8	93.5
Surprisal	–	3.7	11.6	53.7	99.9	–	8.9	22.0	62.4	99.9
Surprisal-Sum	–	3.5	5.9	14.6	92.2	–	7.9	13.7	27.5	95.7
Disagreement	–	3.3	6.2	15.2	90.6	–	7.3	13.0	24.9	92.9
Joint	–	3.7	6.9	14.7	93.8	–	8.2	14.8	25.3	94.9
Teacher	–	**3.0**	**5.3**	13.6	97.9	–	**7.0**	**11.7**	24.3	100.1
SpecEntropy	–	3.2	5.8	11.0	92.7	–	7.6	12.9	21.4	94.7
SpecFlux	–	3.3	5.8	12.5	91.2	–	7.2	12.9	23.4	92.5
LogMel-TV	–	3.2	5.6	**10.5**	**87.9**	–	7.4	12.6	**21.0**	**91.7**

A large score corresponds to greater cumulative movement of the utterance trajectory in log-Mel feature space and therefore receives higher selection priority. Unlike Accumulated Spectral Flux, which operates on L2-normalized linear-frequency magnitude spectra with an L2 frame distance, Log-Mel Temporal Variation measures frame-to-frame changes after Mel-frequency integration and logarithmic compression using an L1 distance. The two criteria are complementary representations of the same general acoustic-variation hypothesis.

4 Experiments

4.1 Experimental Setup

Dataset and pruning protocol. We conduct all experiments on LibriSpeech(Panayotov et al., 2015). The complete training set, containing approximately 281k utterances, is treated as the 100% data reference. We evaluate retention ratios of 30%, 10%, 5%, and 1%. For each selected subset, we train the same downstream ASR model from scratch using an identical training and decoding recipe.

ASR model and evaluation. We use a hybrid CTC/Attention Conformer(Watanabe et al., 2018) as the downstream ASR model. Recognition performance is measured by word error rate (WER) on the LibriSpeech test-clean and test-other evaluation sets, with lower WER indicating better performance.

Pruning implementation. For Surprisal and Surprisal-Sum, we score the ground-truth transcripts with a pretrained GPT-2 language model (Radford et al., 2019). For Joint, we use openai/whisper-small(Radford et al., 2023) as the audio encoder and intfloat/multilingual-e5-baseWang et al. (2024) as the text encoder. For Teacher, the pretrained openai/whisper-small model serves as the Teacher, while the epoch-2 checkpoint of the full-data Conformer serves as the Student. The model-free acoustic criteria require no learned scoring model.

4.2 Performance Across Different Data Budgets

Table 1 reports the WER of all main pruning criteria under retention ratios ranging from 30% to 1%. At relatively larger retention ratios, the model-based Teacher criterion provides the strongest performance. With 30% of the training utterances, Teacher achieves WERs of 3.0 and 7.0 on test-clean and test-other, respectively. It remains the best criterion at 10% retention with 5.3/11.7 WER. Nevertheless, the model-free acoustic criteria remain highly competitive: at 10%, LogMel-TV reaches 5.6/12.6, while SpecEntropy and SpecFlux both obtain 5.8/12.9.

The relative ranking changes as the data budget becomes more restrictive. At 5% retention, LogMel-TV achieves the best WER of 10.5/21.0, improving over Teacher by 3.1 and 3.3 absolute WER points on test-clean and test-other, respectively. SpecEntropy (11.0/21.4) and SpecFlux (12.5/23.4) also outperform Teacher (13.6/24.3), indicating that the advantage is not limited to a single acoustic criterion. This transition is highlighted in Fig. 2, which compares Random, Teacher, and LogMel-TV from 30% to 5% retention. At 30% and 10%, Teacher and LogMel-TV both outperform Random and remain relatively close to each other. At 5%, however, LogMel-TV separates much more clearly

Table 2: Computational cost of different pruning methods on LibriSpeech. "Model" lists the pretrained model(s) used for score computation. $\mathcal{T}$ denotes the model training time required for scoring, $\mathcal{I}$ denotes score computation through learned-model inference, and $\mathcal{C}$ denotes non-model score computation. Total reports the overall pruning cost.

Method	Model	$\mathcal{T}$	$\mathcal{I}$	$\mathcal{C}$	Total
Random	–	–	–	1 s	1 s
Surprisal	GPT-2	–	0.60 h	–	0.60 h
Surprisal-Sum	GPT-2	–	0.59 h	–	0.59 h
Disagreement	–	3.75 h	2.26 h	–	6.01 h
Joint	Whisper + E5	–	6.19 h	–	6.19 h
Teacher	Whisper	3.75 h	9.98 h	–	13.73 h
SpecEntropy	–	–	–	0.41 h	0.41 h
SpecFlux	–	–	–	0.34 h	0.34 h
LogMel-TV	–	–	–	0.42 h	0.42 h

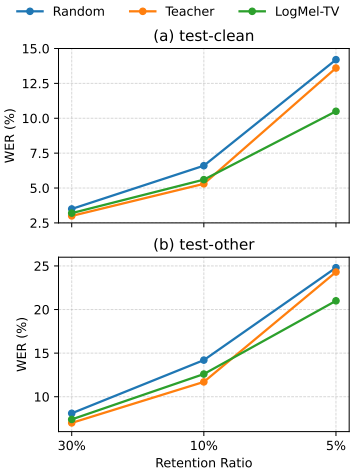


Figure 2: WER of Random, Teacher, and LogMel-TV at different retention ratios.

from both Teacher and Random, suggesting that its acoustic variation signal becomes increasingly useful as the available data budget becomes more severe.

At 1% retention, the absolute WER becomes very high for all methods, making this regime primarily an extreme stress test. Nevertheless, the relative behavior remains consistent with the trend already observed at 5%: LogMel-TV continues to degrade more gracefully than the competing criteria. Taken together, these results indicate that signal-derived acoustic information becomes increasingly valuable under more aggressive data constraints.

4.3 Performance–Efficiency Trade-off

Time Cost Results. We further compare the computational overhead required to construct each pruned subset. Table 2 reports the wall-clock pruning cost. For model-dependent methods, model training and learned-model inference are measured using two NVIDIA A100 GPUs. Non-model score computation, including the three model-free acoustic criteria, is measured using 16 CPU cores. The computational differences are substantial. Teacher incurs the largest overhead, requiring 3.75 hours of Student training and 9.98 hours of learned-model inference, for a total of 13.73 hours. Disagreement and Joint also require more than six hours for model training or representation extraction. In contrast, SpecEntropy, SpecFlux, and LogMel-TV require neither pruning-specific model training nor learned-model inference, and therefore require no GPU resources for subset construction. Using only 16 CPU cores, they complete scoring in 0.34–0.42 hours.

WER–Time Trade-off. Figure 3 jointly visualizes pruning cost and downstream recognition performance. At 10% retention, Teacher achieves the lowest WER, but is located far from the vertical axis because of its 13.73-hour pruning cost. The three model-free acoustic criteria remain close to Teacher along the WER axis while requiring only a small fraction of its scoring time. The advantage becomes stronger at 5% retention. All three model-free acoustic criteria outperform Teacher on both test sets while remaining substantially cheaper, with LogMel-TV providing the lowest WER overall. Thus, the acoustic criteria offer a favorable performance–efficiency trade-off: at moderate pruning they approach the accuracy of the substantially more expensive model-based criterion, while under more aggressive pruning they can simultaneously achieve lower WER and lower pruning cost.

4.4 Analysis of Cumulative Scoring

Several of the strongest criteria in Table 1 use cumulative rather than length-normalized scores. We therefore investigate two questions: whether accumulation itself is important under an utterance-count pruning budget, and whether the gains of the acoustic criteria can be explained by simply favoring samples containing more linguistic content.

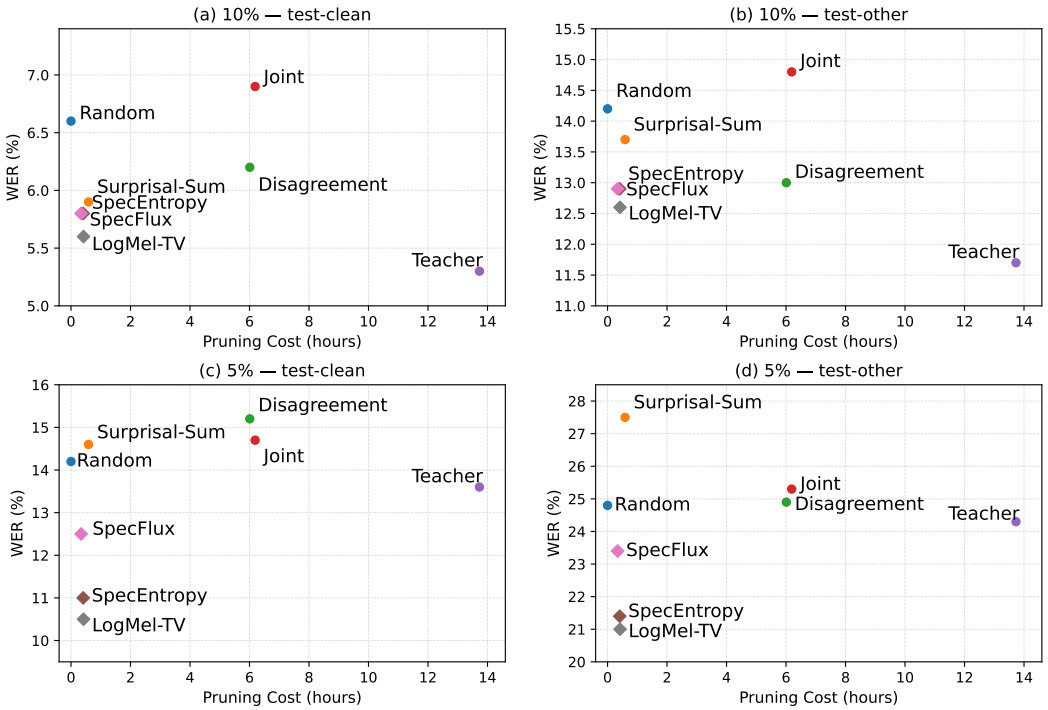


Figure 3: Performance–efficiency trade-off at 10% and 5% retention on LibriSpeech test-clean and test-other. The horizontal axis denotes the total pruning cost from Table 2, and the vertical axis denotes WER. Lower-left positions are preferable, corresponding to both lower pruning cost and lower WER.

Table 3: Analysis of score aggregation and length/content effects. Each WER entry reports test-clean / test-other, and lower is better. (a) compares mean and sum aggregation. (b) compares model-free acoustic criteria with simple linguistic-content quantity baselines.

(a) Mean vs. Sum Aggregation

Criterion	Aggregation	10%	5%
Surprisal	Mean	11.6 / 22.0	53.7 / 62.4
Surprisal	Sum	5.9 / 13.7	14.6 / 27.5
LogMel-TV	Mean	6.8 / 15.1	14.5 /26.4
LogMel-TV	Sum	5.6 / 12.6	10.5 / 21.0

(b) Length/Content Baselines

Method	10%	5%
WordCount	7.3 / 15.7	19.3 / 33.2
PhoneCount	6.0 / 14.1	14.9 / 29.9
SpecEntropy	5.8 / 12.9	11.0 / 21.4
SpecFlux	5.8 / 12.9	12.5 / 23.4
LogMel-TV	**5.6 / 12.6**	**10.5 / 21.0**

Benefits of Cumulative Scoring. Table 3(a) shows a clear advantage for cumulative aggregation. For transcript surprisal, replacing mean token surprisal with its cumulative counterpart reduces WER from 11.6/22.0 to 5.9/13.7 at 10% retention, with an even larger improvement at 5%. The same pattern is observed for LogMel-TV at 10%, where cumulative temporal variation improves WER from 6.8/15.1 to 5.6/12.6.

This behavior is consistent with the pruning budget used in our setting. Because the budget constrains the number of retained utterances rather than their total duration or number of tokens, a mean score measures only the information or variation density within an utterance while discarding how much such information is accumulated over the complete sample. In contrast, cumulative scoring jointly reflects the amount of content and its per-unit informativeness or acoustic variation.

Beyond simple length/content preference. Since the proposed acoustic criteria accumulate frame-level statistics over the complete utterance, one possible explanation for their strong performance is that they simply favor longer or content-richer samples. To examine this alternative

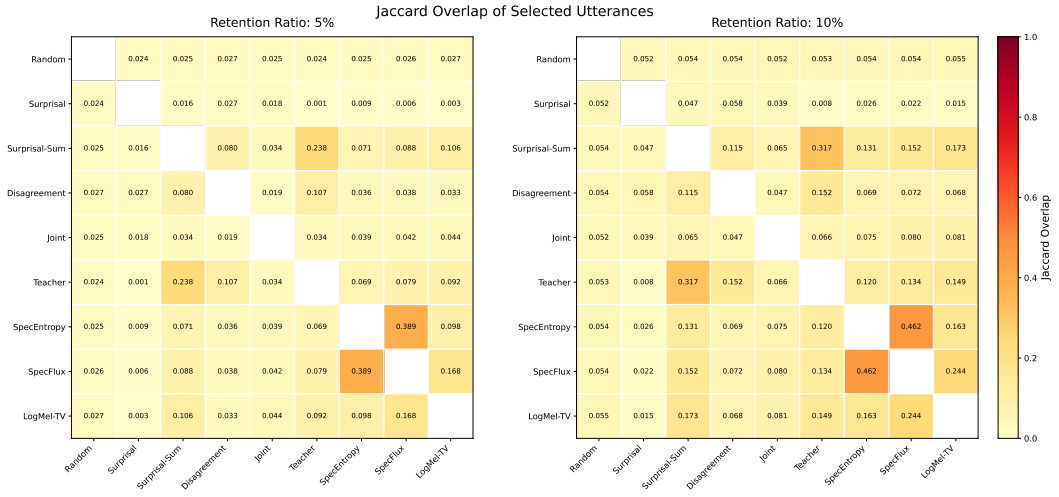


Figure 4: Pairwise overlap between the utterance subsets selected by different pruning methods at 5% (left) and 10% (right) retention on LibriSpeech. Each cell reports the Jaccard similarity between two selected subsets, defined as the size of their intersection divided by the size of their union. Higher values indicate that the two methods retain more similar subsets of training utterances. Diagonal entries are omitted since the self-overlap is trivially one.

explanation, we compare them with WordCount and PhoneCount, two simple quantity-based criteria that similarly assign higher priority to utterances containing more linguistic or phonetic content.

As shown in Table 3(b), both quantity-based baselines are consistently weaker than the model-free acoustic criteria. At 10% retention, PhoneCount yields 6.0/14.1 WER, yet remains inferior to all three acoustic criteria. This advantage becomes more pronounced at 5% retention, where the acoustic criteria clearly outperform both WordCount and PhoneCount. These results suggest that the advantage of the acoustic criteria cannot be attributed solely to their preference for utterances containing more content. They additionally exploit how the speech signal varies within each utterance.

4.5 Selection Behavior and Qualitative Analysis

We further examine whether different pruning criteria retain the same utterances using the Jaccard overlap between their selected subsets. The overlap between Random and the other criteria remains close to these values in Fig. 4, providing a useful reference for interpreting the non-random similarities between pruning methods.

The strongest consistency appears within the acoustic family. SpecEntropy and SpecFlux achieve Jaccard overlaps of 0.389 and 0.462 at 5% and 10% retention, respectively, far above the random reference. This high overlap is consistent with their closely matched performance, particularly at 10%, where both obtain 5.8/12.9 WER. LogMel-TV also shares substantially more samples with the two spectral criteria than the others.

5 Conclusion

We propose a model-free acoustic pruning approach for ASR based on accumulated spectral entropy, spectral flux, and log-Mel temporal variation. These criteria select training utterances directly from speech signals without pruning-specific model training or learned-model inference. Experiments with Conformer on LibriSpeech demonstrate simultaneous improvements in recognition performance and pruning cost over Teacher–Student selection at 5% retention, with Log-Mel Temporal Variation achieving the lowest WER among the evaluated criteria. At larger retention budgets, the acoustic criteria remain competitive while requiring substantially less selection computation. Aggregation ablations further support cumulative scoring under an utterance-count budget. These findings highlight the value of simple acoustic statistics for efficient ASR training-data selection.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and polishing.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed data pruning methods and experimental settings in the main paper to facilitate reproducibility. The code associated with this work will be publicly released.

REFERENCES

Zakaria Aldeneh, Skyler Seto, Maureen de Seyssel, Jie Chi, Zijin Gu, Takuya Higuchi, Jeeweon Jung, Shinji Watanabe, David Grangier, Barry-John Theobald, and Tatiana Likhomanenko. Which data matter? embedding-based data selection for speech recognition. *arXiv preprint arXiv:2603.05819*, 2026. URL https://arxiv.org/abs/2603.05819.

Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. *arXiv preprint arXiv:1906.11829*, 2019.

Yalun Dai, Lingao Xiao, Ivor W. Tsang, and Yang He. Training-free dataset pruning for instance segmentation. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2503.00828.

Hadi M Dolatabadi, Sarah Erfani, and Christopher Leckie. -robustness and beyond: Unleashing efficient adversarial training. In *European conference on computer vision*, pp. 467–483. Springer, 2022.

Xijie Huang, Zechun Liu, Shih-Yang Liu, and Kwang-Ting Cheng. Efficient and robust quantization-aware training via adaptive coreset selection. *arXiv preprint arXiv:2306.07215*, 2023.

Krishnateja Killamsetty, Sivasubramanian Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pp. 5464–5474. PMLR, 2021.

Jian Luo, Jianzong Wang, Ning Cheng, and Jing Xiao. Loss prediction: End-to-end active learning approach for speech recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021. URL https://arxiv.org/abs/2107.04289.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 650–663, 2021.

Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15630–15649. PMLR, 2022. URL https://proceedings.mlr.press/v162/mindermann22a.html.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020.

Hemant Misra, Shajith Ikbal, Hervé Bourlard, and Hynek Hermansky. Spectral entropy based feature for robust ASR. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004. doi: 10.1109/ICASSP.2004.1325955. URL https://publications.idiap.ch/publications/show/628.

Ashish Mittal, Durga Sivasubramanian, Rishabh Iyer, Preethi Jyothi, and Ganesh Ramakrishnan. Partitioned gradient matching-based data subset selection for compute-efficient robust ASR training. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5999–6010. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.443. URL https://aclanthology.org/2022.findings-emnlp.443/.

Robert C. Moore and William Lewis. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pp. 220–224. Association for Computational Linguistics, 2010. URL https://aclanthology.org/P10-2041/.

Juhan Nam, Jorge Herrera, and Kyogu Lee. A deep bag-of-features model for music auto-tagging. *arXiv preprint arXiv:1508.04999*, 2015. URL https://arxiv.org/abs/1508.04999.

Binh-Nguyen Nguyen and Yang He. Swift cross-dataset pruning: Enhancing fine-tuning efficiency in natural language understanding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 726–739, 2025.

Patrik Okanovic, Roger Waleffe, Vasilis Mageirakos, Konstantinos Nikolakakis, Amin Karbasi, Dionysios Kalogerias, Nezihe Merve Gürel, and Theodoros Rekatsinas. Repeated random sampling for minimizing the time-to-accuracy of learning. In *International Conference on Learning Representations*, volume 2024, pp. 24058–24098, 2024.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/ac56f8fe9eea3e4a365f29f0f1957c55-Abstract.html.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Zhenyu Tang, Shaoting Zhang, and Xiaosong Wang. Exploring data redundancy in real-world image classification through data selection. *arXiv preprint arXiv:2306.14113*, 2023.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*, 2018.

Max Welling. Herding dynamical weights to learn. In *Proceedings of the 26th annual international conference on machine learning*, pp. 1121–1128, 2009.

Ryan Whetten, Titouan Parcollet, Marco Dinarelli, and Yannick Estève. A study of data selection strategies for pre-training self-supervised speech models. In *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026. URL https://arxiv.org/abs/2601.20896.

Xiaobo Xia, Jiale Liu, Jun Yu, Xu Shen, Bo Han, and Tongliang Liu. Moderate coreset: A universal method of data selection for real-world data-efficient deep learning. In *The Eleventh International Conference on Learning Representations*, 2022.

Qiao Xiao, Pingchuan Ma, Adriana Fernandez-Lopez, Boqian Wu, Lu Yin, Stavros Petridis, Mykola Pechenizkiy, Maja Pantic, Decebal Constantin Mocanu, and Shiwei Liu. Dynamic data pruning for automatic speech recognition. In *Interspeech 2024*, pp. 4488–4492, 2024. doi: 10.21437/Interspeech.2024-1330. URL https://www.isca-archive.org/interspeech_2024/xiao24b_interspeech.html.

Xilie Xu, Jingfeng Zhang, Feng Liu, Masashi Sugiyama, and Mohan S Kankanhalli. Efficient adversarial contrastive learning via robustness-aware coreset selection. *Advances in Neural Information Processing Systems*, 36:75798–75825, 2023.