



BR-ASR: Efficient and Scalable Bias Retrieval Framework for Contextual Biasing ASR in Speech LLM

Xun Gong¹, Anqi Lv², Zhiming Wang², Huijia Zhu², Yanmin Qian^{1*}

¹Auditory Cognition and Computational Acoustics Lab, MoE Key Lab of AI, AI Institute School of Computer Science, Shanghai Jiao Tong University, Shanghai, China

²AntGroup, Shanghai, China

{gongxun, yanminqian}@sjtu.edu.cn, {lyuanqi.laq, zhiming.wzm, huijia.zhj}@antgroup.com

Abstract

While speech large language models (SpeechLLMs) have advanced standard automatic speech recognition (ASR), contextual biasing for named entities and rare words remains challenging, especially at scale. To address this, we propose **BR-ASR**: a **B**ias **R**etrieval framework for large-scale contextual biasing (up to 200k entries) via two innovations: (1) speech-and-bias contrastive learning to retrieve semantically relevant candidates; (2) dynamic curriculum learning that mitigates homophone confusion which negatively impacts the final performance. The is a general framework that allows seamless integration of the retrieved candidates into diverse ASR systems without fine-tuning. Experiments on LibriSpeech test-clean/-other achieve state-of-the-art (SOTA) biased word error rates (B-WER) of 2.8%/7.1% with 2000 bias words, delivering 45% relative improvement over prior methods. BR-ASR also demonstrates high scalability: when expanding the bias list to 200k where traditional methods generally fail, it induces only 0.3/2.9% absolute WER/B-WER degradation with a **99.99%** pruning rate and only 20ms latency per query on test-other.

Index Terms: contextual bias, speech retrieval, contrastive learning, speech large language model, speech recognition

1. Introduction

The integration of speech processing into large language models (SpeechLLMs) has advanced speech tasks like automatic speech recognition (ASR), achieving state-of-the-art performance in general domains [1], making them more applicable in real-world scenarios such as voice assistants. However, challenges arise in accurately recognizing and understanding bias words (also known as hotwords), which are rare terms such as names, locations, or uncommon slang [2, 3, 4, 5]. To address this issue, many contextual biasing ASR systems are developed to improve the bias word recognition accuracy without dropping the overall ASR performance [6, 3, 7].

A straightforward approach is to post-process ASR hypotheses with bias words by leveraging text-only LLMs' linguistic reasoning ability [8, 9]. However, these text-only approaches inherently lack speech modality awareness (e.g., accent, prosody), resulting in suboptimal performance compared to speech-grounded methods [10]. This limitation has driven recent efforts to develop prompt-based contextual biasing through multimodal SpeechLLMs [11, 10, 6, 12, 13, 14], where acoustic-textual alignment enhances robustness and recognition accuracy. In these methods, the prompt typically comprises a bias word list or domain-specific instructions. However, three critical limitations persist: (1) Excessive prompt

length induces LLM hallucinations, leading to unreliable ASR outputs; (2) Increased computational complexity from lengthy prompts degrades inference efficiency. (3) Bias list is limited to dozens of entries, severely limiting the scalability.

To scale contextual biasing in ASR, recent approaches adopt retrieval-augmented generation (RAG) paradigms from LLMs [15] integrating external knowledge database. RAG Methods like [16, 17, 18] use LLMs for ASR hypotheses post-processing. However, two more challenges in scalability and efficiency exist: (1) Domain-specific knowledge bases lack fine-grained alignment with exact bias words, limiting the bias recognition accuracy. (2) Retrieval mechanisms rely on rough cross-modal similarity without deep speech-text modality alignment.

Current fine-grained biasing methods exploit localized acoustic or semantic patterns. Phoneme-RAG [19] employs multi-stage pipelines, where entities are detected at the first stage and then retrieved from the whole database and fed to the LLM. CTC-filter [7] utilizes CTC N-best hypotheses to filter relevant bias words. However, Phoneme-RAG's two-stage inference incurs significant latency with limited performance improvement. CTC-filter suffers from irreversible errors when correct candidates are absent from CTC N-best hypotheses. While achieving better accuracy, both methods scale poorly beyond 10k entries and lack dynamic database integration, fundamentally limiting their industrial applicability.

Recent attention-based advancements focus on frame-level features, and extend the bias number to 40k–100k entries through Dual-NAM [20], and VQ-RAG [21]. However, the performance still degrades linearly beyond 10k entries with excessive computational overhead. This underscores the unresolved trade-off between bias list size, accuracy, and inference speed in large-scale contextual biasing ASR.

To bridge these gaps, we propose **BR-ASR**: an efficient and scalable **B**ias **R**etrieval framework that fundamentally mitigates scalability bottlenecks for contextual ASR. Our approach begins with a contrastive learning mechanism that optimizes speech-bias matching via two different bias encoding strategies, enabling effective retrieval between speech inputs and large-scale bias entries. Besides, speech encoding and bias encoding procedures are inherited from SpeechLLMs, simplifying the training procedure. Building on this foundation, we design a dynamic curriculum scheduler that automatically modulates hard negative sampling ratios during training, specifically homophone confusion, a critical bottleneck in large-vocabulary biasing. The proposed BR-ASR framework is evaluated on the LibriSpeech benchmark with Rare5k [22]. The proposed framework achieves 93%/91% retrieval accuracy on test-clean/-other subsets and delivers state-of-the-art B-WER of 2.8%/7.1% over all previous methods. As the number of bias words scales up to 200k, it preserves a stable performance

*Corresponding author

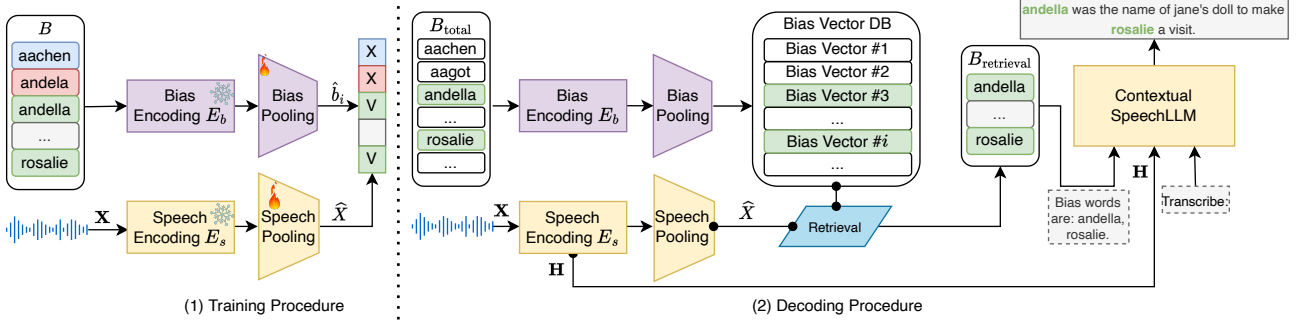


Figure 1: The proposed Bias Retrieval Framework with training (1) and decoding (2) procedures for Contextual SpeechLLM.

with only 0.3%/2.9% absolute WER/B-WER degradation on the test-other subset, resolving the long-tail coverage issues. With the help of FAISS [23], the retrieval latency is only 20ms when querying 200k entries on GPU. Crucially, our framework demonstrates strong generalization: when integrated with third-party ASR systems, it consistently improves rare-word recognition without architectural modifications.

2. Methodology

2.1. Revisiting Contextual Biasing in SpeechLLM-based ASR Systems

The contextual ASR (C-ASR) can be formally defined through three core components: (1) the input audio spectrogram $\mathbf{X} \in \mathbb{R}^{M \times T}$, where M denotes mel-frequency bins and T temporal frames, (2) a predefined bias list $\mathcal{B} = \{b_1, b_2, \dots, b_N\}$ containing N domain-specific terms or rare entities (where the oracle bias words is defined as $\mathcal{B}_{\text{sentence}}$, s.t. $|\mathcal{B}_{\text{sentence}}| \leq N$), and (3) the target transcription \mathbf{y} . The objective is modeled as an autoregressive conditional probability:

$$P(y_{i+1} | \mathbf{X}, \mathbf{y}_{1:i}, \mathcal{B}) = \text{C-ASR}(\phi(\mathbf{X}), \mathbf{y}_{1:i}, \psi(\mathcal{B})), \quad (1)$$

where SpeechLLM implementations involve two critical processes: (1) Speech encoding $\phi(\cdot)$: Audio inputs are encoded through $\phi(\cdot)$ into latent representations \mathbf{H} , which serve as speech latent features for the SpeechLLM. (2) Bias integration: The bias list \mathcal{B} is incorporated via $\psi(\cdot)$, implemented either as textual prompts (e.g., “Related bias words are b_1, b_2, \dots ”). [10, 24, 6] or through dedicated bias encoders [25, 26].

2.2. Bias Retrieval Framework

2.2.1. Contrastive Learning for Speech-and-Bias Alignment

Direct similarity computation between textual bias words and raw speech signals proves ineffective due to modality discrepancies and granularity mismatches. To address this, we redesign contrastive language-audio pretraining (CLAP) [27] for speech-and-bias alignment, as illustrated in Fig. 1 (1).

Our modified training procedure introduces two encoding network, speech encoding network E_s , and bias encoding network E_b , the pre-training procedure can be formulated as:

$$\hat{\mathbf{X}} = \text{Pool} \cdot E_s(\mathbf{X}), \quad \hat{\mathbf{b}} = E_b(b), \quad (2)$$

$$\mathcal{L} = 0.5 * (l_a(S) + l_b(S)), \text{ where } S = \tau(\hat{\mathbf{X}} \cdot \hat{\mathbf{b}}^T), \quad (3)$$

where $\hat{\mathbf{X}} \in \mathbb{R}^d, \hat{\mathbf{b}} \in \mathbb{R}^d$ are encoded speech/bias features of dimension d , and $\text{Pool}(\cdot)$ is a special network to make speech features the same length as the bias features. For the objective loss, S is the dot-product similarity score, τ is a learnable tem-

perature and l_k is the sum of logits along text and audio axes respectively. Following CLAP [27], let N_{neg} be the number of negative samples during training, we randomly select N_{neg} bias words from the bias database $\mathcal{B}_{\text{total}} \setminus \mathcal{B}_{\text{sentence}}$, where $\mathcal{B}_{\text{sentence}}$ is the oracle bias list for each sentence.

2.2.2. Frozen Encoding Priors

To prevent catastrophic forgetting and ensure feature consistency, we fully freeze speech/bias encoding network E_s, E_b shown in Fig. 1 (1).

Frozen Encoding Priors: As a well-trained speech encoder is available in SpeechLLMs, we reuse the speech encoder $E_s(\cdot) = \phi(\cdot)$, and thereby the dimension mismatch between speech latent and bias latent can be simplified into two parameter-efficient ways for different bias encoding network:

- *AcousticBias*: The bias word is firstly synthesized to audio, then the same encoding network is applied as $E_b(b) = E_s(\text{TTS}(b))$. This approach simplifies cross-modal alignment by projecting both speech inputs and bias terms into the same acoustic space.
- *TextualBias*: We also directly employ the same LLM backend to extract textual bias features, i.e. $E_b(b) = \text{LLM}(b)$. This preserves linguistic semantics while maintaining model interoperability. For this setup, $E_s(\cdot) = \text{Projection} \cdot \text{Speech-Encoder}(\cdot)$.

Pooling Strategies for speech representation To address length mismatch between speech latent features and bias latent features, we implement three strategies to pool the frame-level speech features \mathbf{H} from $D \times T'$ to the pooled feature $\hat{\mathbf{X}}$:

- *Adaptive Average Pooling*: The same pooling as CLAP [27].
- *Attention Pooling*: Let d be the dimension of attention calculations, the output is $\hat{\mathbf{X}} = \text{Linear} \cdot \sum_{T'} (\text{Attention}(\mathbf{H}))$.

2.2.3. Dynamic Curriculum Learning for Homophone Debias

To mitigate homophone confusion (e.g., “there” vs. “their”), we develop a phoneme-aware dynamic curriculum learning (DCL) mechanism [28, 29, 30]. Firstly, we use grapheme-to-phoneme (G2P) to generate phoneme sequence and and construct homophone set \mathcal{H}_i for bias word b_i :

$$\mathcal{H}_i = \{h_j | \text{Lev}(p_i, p_j) \leq \theta, h_j \in \mathcal{B} \setminus \{b_i\}\}, \quad (4)$$

where Lev is Levenshtein edit distance and $\theta = 2$ is the phoneme mismatch threshold. Then we dynamically adjust the homophone sampling ratio through sigmoidal scheduling:

$$\alpha_n = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min}) \cdot \left(\frac{2}{1 + e^{-\gamma n}} - 1 \right), \quad (5)$$

where n denotes the training steps, and γ is a controllable parameter. To further disperse homophone in training, we apply an additional regularization term in the objective function:

$$\mathcal{L}_{\text{total}} = \mathcal{L} + \lambda \sum_{h_j \in \mathcal{H}_i} \|b_i - h_j\|_2^2, \quad (6)$$

where b_i is the bias word embedding and h_j is word embedding from one of the homophones of b_i .

3. Experiments

3.1. Experimental Setup

3.1.1. Dataset and Metrics

Experiments were conducted using LibriSpeech [31] for both training and evaluation, where bias words are marked by Rare5k [22]. Rare5k identifies the top 5k most frequent words as common, with the remaining 209.2k words classified as rare in the LibriSpeech train-960h set.

Following the same setup in [22], we evaluate word error rate (WER) (%) and biased WER (B-WER) (%), where B-WER is calculated over bias words on test-clean/-other subsets.

We further consider two extra evaluation dimensions to directly evaluate the retrieval performance by special recall variants*. (1) **Bias Retrieval Recall**: the recall rate (%) (\uparrow) of bias list over all sentences, i.e. $\text{Recall}_B = \frac{\sum |\mathcal{B}_{\text{sentence}}|}{\sum |\mathcal{B}_{\text{retrieval}}|}$, where $\mathcal{B}_{\text{retrieval}}$ denotes the retrieved bias list. (2) **Homophone Retrieval Recall**: the recall rate (%) (\downarrow) of all homophones over all sentences, i.e. $\text{Recall}_H = \frac{\sum |\mathcal{H}_{\text{sentence}}|^\dagger}{\sum |\mathcal{B}_{\text{retrieval}}|}$.

3.1.2. Model setups and hyper-parameter selection

Two different SpeechLLMs are selected as the contextual ASR backend, Qwen-Audio [32] and SLAM-ASR [33]. For Contextual Qwen-Audio, we follow two different contextual methods, i.e., bias-prompt (Prompt-QwenAudio) and bias-encoder (BiasEncoder-QwenAudio) of [11]. The SpeechLLM backend has been changed to the unchat version[‡]. The Prompt-QwenAudio is trained using the bias size between 20~80. The BiasEncoder-QwenAudio is trained using the bias size between 50~1000, and the bias encoder is initialized from gte-large-en-v1.5[§]. For the Contextual SLAM-ASR, we reuse the pre-trained Prompt-SLAM-ASR [7], where the training bias size is 5 on average.

For the proposed Bias Retrieval, hyper-parameters are set as follows. For AcousticBias defined in Section 2.2.2, we firstly synthesized all bias words from the database $\mathcal{B}_{\text{total}}$ using edge-tts[¶] with the default voice “en-US-AnaNeural”. Then both audio and bias encodings are fed to the same speech encoder used in SpeechLLMs. For TextualBias defined in Section 2.2.2, the bias encoding reuses the textual LLM part from SpeechLLMs, and the audio encoding is done by the speech encoder with projection. Note that there is a dimension mismatch between these encodings, so the same projection inherited from SpeechLLMs is reused. For the contrastive learning

*Recall@X denotes the average retrieval number $\bar{N}_{\text{retrieval}}$ when the Recall = X% over dataset. Recall#X denotes the recall rate (%) when $|\mathcal{B}_{\text{retrieval}}| = X$ for each sentence.

[†] $\mathcal{H}_{\text{sentence}} = \{h|h \in \bigcup_i \mathcal{H}_i, \forall i \in \mathcal{B}_{\text{sentence}}\}$

[‡]<https://huggingface.co/Qwen/Qwen-Audio>

[§]<https://huggingface.co/Alibaba-NLP/gte-large-en-v1.5>

[¶]<https://github.com/rany2/edge-tts>

proposed in Section 2.2.1, the temperature τ is initialized to 0.007 and clipped by at most 100 in Eq. 3. During training, we use AdamW with a maximum learning rate of $1e^{-4}$ warmup ratio of 0.05 and train the network for 10 epochs with 960h data. As for the dynamic curriculum learning in Section 2.2.3, we set the homophone distance threshold θ to 2, and the parameters as $\alpha_{\text{min}} = 0.01, \alpha_{\text{max}} = 0.5, \gamma = 0.05$. The regularization loss parameter λ in Eq. 6 is set to 0.1.

Finally, for the scalable retrieval, we use FAISS [23], where vectors are pre-normalized and then retrieved with METRIC_INNER_PRODUCT. The retrieval threshold is set to 50, 50, and 10 for Prompt-QwenAudio, BiasEncoder-QwenAudio and Prompt-SLAM-ASR, respectively.

3.2. Main Results of Bias Retrieval Method

Table 1 compares the performance of contextual ASR models on LibriSpeech test sets. We firstly explore the performance on QwenAudio variants. The baseline QwenAudio achieves WER=2.0%/4.2% and B-WER=8.4%/18.4% on test-clean/-other, respectively. Its fine-tuned version obtains WER=1.7%/3.8% and B-WER=8.3%/18.5%, indicating that pure domain adaptation fails to improve bias recognition accuracy. We reproduce two implementations based on [11], both surpassing the original CB-QwenAudio. Results show that Prompt-QwenAudio suffers catastrophic failure (at $N \geq 100$) due to LLM hallucinations, and unreliable results for $N \geq 1000$ due to exceeded context limits. In contrast, BiasEncoder-QwenAudio obtains better results when $N = 1000, 2000$.

With the proposed Bias Retrieval framework, both achieve systematic improvements: Prompt-QwenAudio+BR reduces B-WER by over 80% absolutely when $N=500$ (6.3% vs 32.1%), and continues working well when $N=1000/2000$. BiasEncoder-QwenAudio+Bias-Retrieval keeps the B-WER below 10.2% for nearly all N values, which is attributed to its ability to maintain better stability through the secondary text processing that effectively disambiguates homophones.

Compared to conventional approaches, previous SOTA, USTR-CT exhibits a B-WER increase from 4.4% to 7.6%. In contrast, our method demonstrates superior robustness while maintaining a stable B-WER (7.1% for Prompt+BR and 10.2% for BiasEncoder+BR), by modeling speech-and-bias correlation. This also represents 17.1% relative improvement over USTR-CT’s best baseline performance when $N=2000$. Our framework ultimately achieves breakthrough performance with B-WER=2.8%/7.1% when $N=2000$, establishing a new SOTA for large-scale contextual ASR.

In addition the results on the SLAM-ASR system also validate our framework’s generalization capability, showing similar trends in preserving the B-WER performance.

3.3. Ablation Studies

Based on the ablation study results in Table 2, we systematically analyze key designs of the bias retrieval framework:

Speech Pooling Strategy: It shows that AcousticBias with AdaAvgPool does not work effectively as $\text{Recall}_H@99 > 50$. In contrast, AttnPool provides a better pooling, focusing on the semantic information and increasing the retrieval efficiency: $\text{Recall}_B@99$ reduced greatly from 54.1 to 38.7. This also implies that almost all bias words can be successfully retrieved within top-50 ($\text{Recall}_B\#50$). With AttnPool, the B-WER improves by 32% relatively.

Dynamic Curriculum Learning: Although $\text{Recall}_H\#50$ is high, we found that there is a huge gap between $N =$

Table 1: Performance of the contextual ASR baselines and our proposed **Bias Retrieval** method on LibriSpeech test-clean/-other sets.

Contextual ASR Model	$N = \mathcal{B}_{\text{sentence}} $		$N = 100$		$N = 500$		$N = 1000$		$N = 2000$	
	WER	B-WER	WER	B-WER	WER	B-WER	WER	B-WER	WER	B-WER
DB-NNLM [22]	-	-	2.0/5.9	5.7/14.1	2.1/6.1	6.2/15.1	2.1/6.4	6.7/17.2	2.3/6.6	7.3/18.9
USTR-CT [34]	-	-	2.1/5.4	2.0/4.4	2.1/5.6	2.2/5.6	2.2/5.8	2.5/6.3	2.2/5.8	3.0/7.6
CB-QwenAudio [11]	-	-	1.6/3.8	5.5/13.5	1.9/3.9	6.0/14.2	-	-	-	-
Prompt-QwenAudio	1.1/2.6	1.2/3.7	1.3/3.5	2.4/6.4	22.4/27.8	32.1/39.1	X	X	X	X
+ Bias Retrieval	1.1/2.6	1.2/3.7	1.2/2.7	2.3/6.1	1.2/2.7	2.4/6.3	1.2/2.7	2.6/6.6	1.2/2.8	2.8/7.1
BiasEncoder-QwenAudio	1.4/3.3	3.8/9.0	1.5/3.6	5.0/13.0	1.6/3.7	5.6/13.8	1.7/3.9	6.2/14.7	1.8/4.1	7.4/18.1
+ Bias Retrieval	1.4/3.3	3.8/9.0	1.5/3.4	4.0/9.5	1.5/3.5	4.2/9.6	1.5/3.6	4.3/9.8	1.5/3.6	4.3/9.9
Prompt-SLAM-ASR [7]	1.1/2.7	2.8/6.0	7.4/17.9	24.8/44.1	X	X	X	X	X	X
+ CTC-Filter [7]	1.1/2.7	2.8/6.0	1.3/2.7	3.7/8.0	1.3/3.0	3.9/9.0	1.3/3.0	4.2/9.3	1.4/3.2	4.4/10.0
+ Bias Retrieval	1.1/2.7	2.8/6.0	1.3/2.7	3.8/8.3	1.3/3.0	4.0/8.7	1.3/3.1	4.1/9.0	1.4/3.1	4.2/9.3

Table 2: Ablation Studies on LibriSpeech test-other set with $N = 2000$ (2k) of Prompt-QwenAudio as Contextual ASR system.

Bias Modality	Speech Pooling	Curriculum Learning	Recall _B @99 (↑)	Recall _B #50 (↑)	Recall _H #50 (↓)	WER	B-WER
Acoustic	AdaAvgPool	✗	54.1	98.1%	98.7%	3.3	12.4
	AttnPool	✗	32.8	99.9%	99.9%	3.0	8.6
	AttnPool	✓	42.2	99.7%	69.3%	2.9	7.4
Textual	AdaAvgPool	✗	58.5	97.3%	94.9%	3.1	10.3
	AttnPool	✗	38.7	99.5%	91.3%	3.0	7.9
	AttnPool	✓	47.3	99.1%	58.4%	2.8	7.1

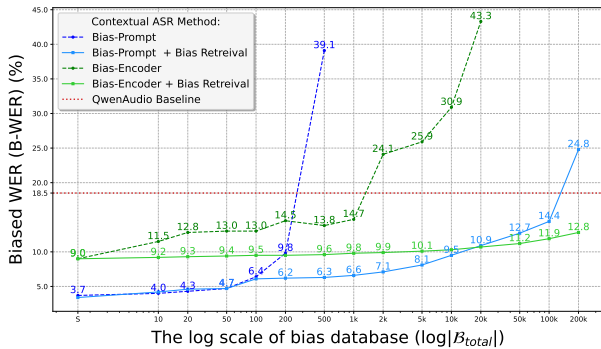


Figure 2: The B-WER (%) performance across different bias database ($\mathcal{B}_{\text{total}}$) scales from 10 to 200k entries. Specially, $S = |\mathcal{B}_{\text{sentence}}|$.

$|\mathcal{B}_{\text{sentence}}|$ (In Table 1, B-WER=3.7% on test-other) and $N = |\mathcal{B}_{\text{retrieval-top-50}}|$. We observed that homophones are mistakenly recognized (i.e., homophones confusion), as Recall_H#50 is also high for both AdaAvgPool and AttnPool. The proposed dynamic curriculum learning method greatly suppresses homophone errors. When activated, the AcousticBias reduces Recall_H#50 from 99.9% to 69.3%, with another 23% relative reduction on B-WER on test-other.

Bias Encoding Modality: To further mitigate homophone errors during retrieval, we adopt TextualBias, as textual inputs avoid the domain mismatch between synthesized and real speech, and are more robust to OOV words. Compared to AcousticBias, TextualBias achieves competitive retrieval performance and lowers the homophone recall rate (Recall_H#50 reduced by 10%), with a 0.3% absolute reduction in B-WER.

3.4. Scalability Analysis

For large-scale bias database up to 200k entries, Fig. 2 shows that the B-WER degradation is manageable. The BiasEn-

coder method shows relatively stable performance, where B-WER only degrades by 2.9% (from 9.9% to 12.8%) as $|\mathcal{B}_{\text{total}}|$ expands from 2k to 200k (i.e., scaling up by 100x), as the Prompt method performance deteriorates faster. The framework achieves real-time efficiency through FAISS-accelerated retrieval: on a single-thread CPU (Intel Xeon Gold 5318Y), it processes 200k entries with 4096-dimensional vectors in 500ms under 7GB RAM, while GPU acceleration (NVIDIA RTX 4090) reduces the latency to 20ms within 8GB VRAM.

4. Conclusion

In this paper, we present BR-ASR: a bias retrieval framework that fundamentally resolves the scalability bottleneck in contextual ASR. Our proposed method has 4 innovations: (1) Cross-modal contrastive learning aligns speech inputs with textual bias entries via pretrained speech and bias encoders, achieving 93% retrieval accuracy while pruning 99.99% of irrelevant candidates at 200k-scale where the performance degradation is only 0.3%/2.9% in terms of WER/B-WER for test-other set. (2) Dynamic curriculum learning suppresses homophone errors by 30.6% absolute recall through progressive hard sample scheduling. Evaluations on LibriSpeech test-clean/-other establish new state-of-the-art performance: 1.2%/2.8% WER and 2.8%/7.1% B-WER when $N=2000$. (3) It demonstrates industrial-grade efficiency with 20ms latency when querying 200k candidates. (4) BR-ASR enables plug-and-play integration with diverse systems, by decoupling bias retrieval from ASR decoding.

5. Acknowledgements

This work was supported in part by China NSFC projects under Grants 62122050 and 62071288, in part by Shanghai Municipal Science and Technology Commission Project under Grant 2021SHZDZX0102, and in part by Ant Group and Ant Group Research Intern Program.

6. References

- [1] S. Ji, Y. Chen, M. Fang, J. Zuo, J. Lu, H. Wang, Z. Jiang, L. Zhou, S. Liu, X. Cheng *et al.*, “WavChat: A survey of spoken dialogue models,” *arXiv preprint arXiv:2411.13577*, 2024.
- [2] G. Pundak, T. N. Sainath, R. Prabhavalkar, A. Kannan, and D. Zhao, “Deep context: End-to-end contextual speech recognition,” in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 418–425.
- [3] X. Yang, W. Kang, Z. Yao, Y. Yang, L. Guo, F. Kuang, L. Lin, and D. Povey, “PromptASR for contextualized ASR with controllable style,” *arXiv preprint arXiv:2309.07414*, 2023.
- [4] W. Wang, X. Gong, H. Shao, D. Yang, and Y. Qian, “Text only domain adaptation with phoneme guided data splicing for end-to-end speech recognition,” in *Interspeech 2023*, 2023, pp. 3347–3351.
- [5] A. Salemi, S. Mysore, M. Bendersky, and H. Zamani, “LaMP: When large language models meet personalization,” *arXiv preprint arXiv:2304.11406*, 2023.
- [6] Z. Chen, H. Huang, A. Andrusenko, O. Hrinchuk, K. C. Puvvada, J. Li, S. Ghosh, J. Balam, and B. Ginsburg, “SALM: Speech-augmented language model with in-context learning for speech recognition and translation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 13 521–13 525.
- [7] G. Yang, Z. Ma, Z. Gao, S. Zhang, and X. Chen, “CTC-assisted LLM-based contextual ASR,” in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 126–131.
- [8] G. Song, Z. Wu, G. Pundak, A. Chandorkar, K. Joshi, X. Velez, D. Caseiro, B. Haynor, W. Wang, N. Siddhartha *et al.*, “Contextual spelling correction with large language models,” in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8.
- [9] Y. Asano, S. Hassan, P. Sharma, A. Sicilia, K. Atwell, D. Litman, and M. Alikhani, “Contextual ASR error handling with LLMs augmentation for goal-oriented conversational AI,” *arXiv preprint arXiv:2501.06129*, 2025.
- [10] E. Lakomkin, C. Wu, Y. Fathullah, O. Kalinli, M. L. Seltzer, and C. Fuegen, “End-to-end speech recognition contextualization with large language models,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 406–12 410.
- [11] X. Gong, A. Lv, Z. Wang, and Y. Qian, “Contextual biasing speech recognition in speech-enhanced large language model,” *Proc. Interspeech*, pp. 257–261, 2024.
- [12] H. Hou, X. Gong, W. Zhang, and Y. Qian, “Ranking and selection of bias words for contextual bias speech recognition,” *Proc. Interspeech*, 2025.
- [13] Y. Li, Y. Wu, J. Li, and S. Liu, “Prompting large language models for zero-shot domain adaptation in speech recognition,” in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8.
- [14] X. Gong, A. Lv, Z. Wang, Z. Huijia, and Y. Qian, “Br-asr: Efficient and scalable bias retrieval framework for contextual biasing asr in speech llm,” *Proc. Interspeech*, 2025.
- [15] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, “Retrieval-augmented generation for large language models: A survey,” *arXiv preprint arXiv:2312.10997*, 2023.
- [16] P. Mathur, Z. Liu, K. Li, Y. Ma, G. Karen, Z. Ahmed, D. Manocha, and X. Zhang, “DOC-RAG: ASR language model personalization with domain-distributed co-occurrence retrieval augmentation,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 5132–5139.
- [17] S. Li, H. Shang, D. Wei, J. Guo, Z. Li, X. He, M. Zhang, and H. Yang, “LA-RAG: Enhancing LLM-based ASR accuracy with retrieval-augmented generation,” *arXiv preprint arXiv:2409.08597*, 2024.
- [18] A. Robatian, M. Hajipour, M. R. Peyghan, F. Rajabi, S. Amini, S. Ghaemmaghami, and I. Gholampour, “GEC-RAG: Improving generative error correction via retrieval-augmented generation for automatic speech recognition systems,” *arXiv preprint arXiv:2501.10734*, 2025.
- [19] Z. Lei, X. Na, M. Xu, E. Pusateri, C. Van Gysel, Y. Zhang, S. Han, and Z. Huang, “Contextualization of ASR with LLM using phonetic retrieval-based augmentation,” *arXiv preprint arXiv:2409.15353*, 2024.
- [20] Z. Huang, D. Caseiro, K. Joshi, C. Li, P. Rondon, Z. Wu, P. Zadrzil, and L. Zhou, “Optimizing large-scale context retrieval for end-to-end ASR,” in *Proc. Interspeech*, 2024, pp. 4573–4577.
- [21] N. Flemotomos, R. Hsiao, P. Swietojanski, T. Hori, D. Can, and X. Zhuang, “Optimizing contextual speech recognition using vector quantization for efficient retrieval,” *arXiv preprint arXiv:2411.00664*, 2024.
- [22] D. Le, M. Jain, G. Keren, S. Kim, Y. Shi, J. Mahadeokar, J. Chan, Y. Shanguan, C. Fuegen, O. Kalinli *et al.*, “Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion,” *arXiv preprint arXiv:2104.02194*, 2021.
- [23] J. Johnson, M. Douze, and H. Jégou, “Billion-scale similarity search with GPUs,” *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.
- [24] X. Gong, Y. Wu, J. Li, S. Liu, R. Zhao, X. Chen, and Y. Qian, “Longfnt: Long-form speech recognition with factorized neural transducer,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [25] X. Gong, W. Wang, H. Shao, X. Chen, and Y. Qian, “Factorized aed: Factorized attention-based encoder-decoder for text-only domain adaptive asr,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [26] X. Gong, Y. Wu, J. Li, S. Liu, R. Zhao, X. Chen, and Y. Qian, “Advanced long-content speech recognition with factorized neural transducer,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1803–1815, 2024.
- [27] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, “CLAP learning audio concepts from natural language supervision,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [28] X. Gong, Z. Zhou, and Y. Qian, “Knowledge transfer and distillation from autoregressive to non-autoregressive speech recognition,” in *Proc. Interspeech 2022*, 2022, pp. 2618–2622.
- [29] X. Gong, Z. Chen, Y. Yang, S. Wang, L. Wang, and Y. Qian, “Speaker embedding augmentation with noise distribution matching,” in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021, pp. 1–5.
- [30] Y. Qian, X. Gong, and H. Huang, “Layer-wise fast adaptation for end-to-end multi-accent speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, p. DOI:10.1109/TASLP.2022.3198546, 2022.
- [31] 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)*, 2015, pp. 5206–5210.
- [32] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, “Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models,” *arXiv preprint arXiv:2311.07919*, 2023.
- [33] Z. Ma, G. Yang, Y. Yang, Z. Gao, J. Wang, Z. Du, F. Yu, Q. Chen, S. Zheng, S. Zhang, and X. Chen, “An embarrassingly simple approach for LLM with strong ASR capacity,” *arXiv preprint arXiv:2402.08846*, 2024.
- [34] J. Qiu, L. Huang, B. Li, J. Zhang, L. Lu, and Z. Ma, “Improving large-scale deep biasing with phoneme features and text-only data in streaming transducer,” in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8.