



Ranking and Selection of Bias Words for Contextual Bias Speech Recognition

Haoxiang Hou, Xun Gong, Wangyou Zhang, Wei Wang, Yanmin Qian[†]

¹Auditory Cognition and Computational Acoustics Lab
MoE Key Lab of Artificial Intelligence, AI Institute

Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China,

houhaoxiang0701@sjtu.edu.cn, gongxun@sjtu.edu.cn, wyz-97@sjtu.edu.cn,
wangwei.sjtu@sjtu.edu.cn, yanminqian@sjtu.edu.cn

Abstract

Contextual Automatic Speech Recognition (ASR) systems have made significant advancements. However, contextual ASR models face challenges when dealing with a large number of bias words. This paper focuses on addressing the limitations of contextual ASR models in handling a substantial number of bias words. First, to guide the model to focus on the most important words, we propose a novel network serving as a scorer for bias word ranking and selection. Second, as an example, we explore the use of the proposed scorer in conjunction with the contextual Whisper model. We create a new bias word list using a named-entity recognition (NER) model, which is closer to real-world scenarios. The results on the LibriSpeech dataset with the IS21 bias words list demonstrate that bias word ranking and selection can significantly enhance the model's performance in recognizing bias words, achieving a relative reduction of over 40% in the Biased Word Error Rate.

Index Terms: automatic speech recognition, contextual biasing, Whisper

1. Introduction

In recent years, end-to-end automatic speech recognition (ASR) systems have seen significant advancements [1]. They generally fall into three main categories: connectionist temporal classification (CTC) models, attention-based encoder-decoder models, and transducer-based models, each offering distinct advantages that have driven their widespread adoption in modern ASR tasks [2, 3]. Despite achieving strong benchmark performance, standard ASR systems still face challenges with rare words, proper names, and other low-frequency terms. This issue arises from the long-tail distribution of infrequent terms in the training data, which results in inaccurate transcriptions.

In response to these challenges, contextual biasing techniques, such as shallow fusion and deep fusion, have proven effective in enhancing ASR performance by integrating contextual information into the ASR process. Shallow fusion combines a pre-trained Language Model (LM) with the acoustic model during decoding [4, 5, 6, 7, 8, 9, 10, 11]. In contrast, deep fusion [12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22] trains both the acoustic model and LM jointly, enabling deeper interaction between them during inference. Recent research efforts have delved into integrating large-scale foundation models with contextual ASR techniques, such as [23, 24, 25, 26, 27, 28, 29, 30]. This integration aims to leverage the capabilities of these advanced models to enhance the performance of automatic speech

recognition in contextual scenarios, potentially improving the accuracy of recognizing rare words, domain-specific terms, and handling various language nuances more effectively.

However, when confronted with a vast number of potential bias words (e.g., over 1000), contextual ASR models often become overwhelmed and struggle to manage such a large volume efficiently. Specifically, models built upon large-scale foundation models are extremely sensitive to the quantity of bias words. This sensitivity can be attributed to limitations in the contextual length and computational efficiency. The restricted contextual length restricts the model's ability to process and incorporate a large number of bias words, while computational efficiency constraints impede the model's capacity to handle the exponentially increased complexity that comes with a large number of bias words, ultimately affecting the overall performance of the contextual ASR system.

In this work, in order to guide the model in focusing on the most relevant bias words, we reduce the total number of bias words through a carefully designed ranking and selection process before incorporating them into the ASR system. We introduce a novel scorer network to prioritize and select the most relevant bias words. First, we use a text-to-speech (TTS) model to convert bias words into corresponding bias audio, enabling seamless integration with the speech audio. A pre-trained audio encoder then extracts features from both the speech and bias audios. Cross-attention mechanisms are employed to capture cross-modal relationships, improving the model's ability to associate bias words with the speech content. Next, a convolutional neural network (CNN) extracts local patterns from these features, followed by a global pooling layer to aggregate them into global features. Finally, a softmax layer generates scores for each bias word, with higher scores indicating a stronger likelihood that the word is a true bias word in the reference text. Top-k selection reduces the number of bias words fed into the model, thus significantly improving scalability by enabling the model to focus on the most relevant words.

In experiments using the LibriSpeech dataset and the IS21 deep bias word list, ranking the bias words in ascending order and selecting the top 50 resulted in a remarkable reduction of over 40% in the Biased Word Error Rate (B-WER) compared to using all bias words. Additionally, applying our method to the TCPGen-based Whisper model and the NER bias word list yielded relative B-WER reductions of over 10% and 30%, respectively. These consistent improvements across different models and bias word lists highlight the robustness and adaptability of our approach. Our method not only demonstrates effective performance in various experimental settings but also lays a solid foundation for future advancements in ASR systems and practical applications.

[†]: Corresponding Author

2. Methodology

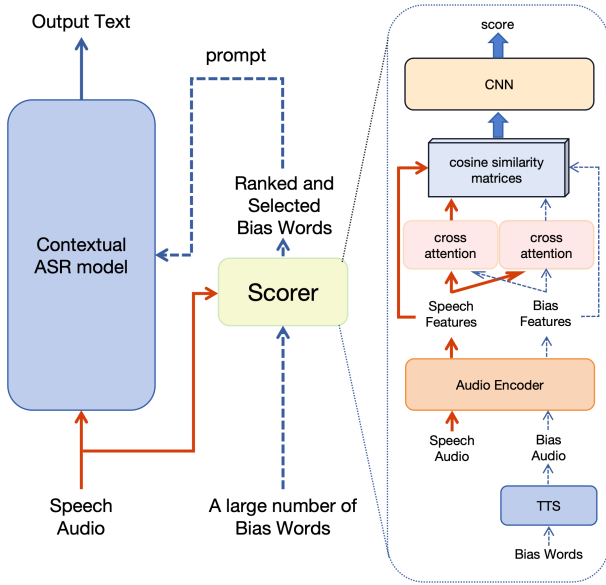


Figure 1: *Model architecture.* We employ a network scorer to rank and select from a large number of bias words. The ranked and selected bias words are then fed into the contextual ASR model. The speech audio is inputted into both the contextual ASR model and the scorer, where it serves as a reference audio.

2.1. Prompt-based Contextual ASR

Contextual ASR models usually involve speech audio and bias words as inputs for generating domain-specific recognition results. Currently, prompt has been a popular approach to inject bias words, capitalizing on the language modeling capabilities of the Whisper decoder or Large Language Model (LLM). In this paper, we also adopt the prompt-based approach and extend it to cope with a wide range of bias word scales, as shown in Figure 1.

We explored two primary methods for providing the prompt: a naive approach and a spoken-style approach. In the naive approach, we simply concatenate the biasing words, separating them with commas and spaces. Conversely, the spoken-style prompt endeavors to arrange the words in a more conversational style. In this study, as per the approach in [31], we use the format “The topic of today’s speech is, ah, {prompt words1, prompt words2, ...}. Okay, then I’ll continue.”

In real-world applications, the number of bias words can easily scale up to over 1000, leading to a largely diluted biasing effect and computational inefficiency in existing prompt-based contextual ASR approaches. To tackle this challenge, we propose leveraging an auxiliary scorer to dynamically rank and select the most relevant bias words, effectively achieving high ASR accuracy while reducing the computation overhead.

2.2. Scorer architecture

The proposed scorer takes as input a predefined list of bias words and the speech audio for ASR, and performs an efficient comparison between the speech and all bias word candidates in the latent space. To map the bias words and speech audio to the same latent space, we utilize a pre-trained TTS model,

which converts each bias word into a spoken audio. Leveraging the unified representation of both speech and bias words in the audio domain, we can then efficiently calculate the pairwise similarity between the speech and bias word audios, which is further used for bias word ranking and selection.

Specifically, a pre-trained audio encoder is utilized to extract features from both speech audio and bias audio separately. Subsequently, a linear projection layer is employed for dimensionality reduction, thus improving the overall efficiency for similarity calculation. We then employ the cross-attention mechanism to extract the cross-modal features between speech audio and bias audio, as shown in Eqn 1. In the first step, we use the speech audio features \mathbf{F}_{speech} as the query, and the bias audio features \mathbf{F}_{bias} as both the key and value in the cross-attention mechanism. This focuses the model on the most relevant bias words to the speech input. Similarly, we reverse the roles in the second cross-attention step, where the \mathbf{F}_{bias} serve as the query, and the \mathbf{F}_{speech} are used as the key and value. This dual cross-attention approach allows the model to learn bidirectional relationships between the speech and bias words, enhancing its ability to link bias words with the speech context.

$$\begin{aligned} \mathbf{F}_{speech \rightarrow bias} &= \text{cross-att}(\mathbf{F}_{speech}, \mathbf{F}_{bias}, \mathbf{F}_{bias}) \\ \mathbf{F}_{bias \rightarrow speech} &= \text{cross-att}(\mathbf{F}_{bias}, \mathbf{F}_{speech}, \mathbf{F}_{speech}) \end{aligned} \quad (1)$$

We compute multiple cosine similarity matrices for multifaceted comparison between speech and bias word features. The first similarity matrix measures the cosine similarity between the speech features and the bias word features, providing a direct comparison of the two modalities. Additionally, we compute similarity matrices for different combinations of features: the cosine similarity between \mathbf{F}_{speech} and $\mathbf{F}_{bias \rightarrow speech}$, the cosine similarity between $\mathbf{F}_{speech \rightarrow bias}$ and \mathbf{F}_{bias} , and the cosine similarity between $\mathbf{F}_{speech \rightarrow bias}$ and $\mathbf{F}_{bias \rightarrow speech}$, as shown in Eqn 2. These matrices allow the system to capture multiple aspects of the relationship between speech and bias word audios.

$$\begin{aligned} S_1 &= \text{sim}(\mathbf{F}_{speech}, \mathbf{F}_{bias}) \\ S_2 &= \text{sim}(\mathbf{F}_{speech}, \mathbf{F}_{bias \rightarrow speech}) \\ S_3 &= \text{sim}(\mathbf{F}_{speech \rightarrow bias}, \mathbf{F}_{bias}) \\ S_4 &= \text{sim}(\mathbf{F}_{speech \rightarrow bias}, \mathbf{F}_{bias \rightarrow speech}) \\ \mathbf{S} &= \text{stack}(S_1, S_2, S_3, S_4) \end{aligned} \quad (2)$$

These similarity matrices are then stacked and fed into a series of CNN layers to extract a powerful feature representation that captures both local and global dependencies between speech and bias words. Finally, to obtain the global features for a final prediction, we use a global pool layer to merge all features from the CNN layers into a set of global features. These global features are then passed through a linear layer followed by a softmax layer to compute the final scores for each bias word relative to the entire speech input. The higher the score, the more likely that the bias word corresponds to a true bias word present in the reference text of the speech. This ranking mechanism allows the system to determine the most relevant bias word, providing a confidence score for whether a specific term or entity is present in the speech. Finally, we utilize the confidence scores to rank the bias words and select the top k words with the highest scores.

To further refine the model and enhance its robustness, we generate more challenging negative samples compared to those randomly selected[32]. Specifically, we base this generation on

the edit distance, choosing the bias word with the minimum edit distance as the negative sample. The edit distance-based approach creates negative samples that are more difficult for the model to distinguish from the true bias words, forcing the model to learn better feature representations and thus improving its overall performance.

3. Experimental Setups

3.1. Datasets

We use the LibriSpeech [33] dataset, a widely used benchmark for ASR, to validate our proposed approach. It consists of approximately 1,000 hours of English speech from audiobooks, divided into clean and noisy subsets for both training and evaluation. In this work, due to the relatively small size of the network, we use train-clean-100 as the training dataset, dev-clean for validation, and test-clean and test-other for testing.

3.2. Biasing list

3.2.1. IS21 deep bias words list

We first follow the IS21 deep bias word list approach[34], where the entire vocabulary is initially defined by the word frequency distribution. Specifically, it removes the 5,000 most common words. The remaining words, regarded as rare, are used to compile the biasing list. In total, there are 209,291 rare words. Additionally, it introduces distractors by randomly sampling from the set of rare words in the training vocabulary, with different numbers of distractors $N = \{100, 500, 1000, 2000\}$ used for each test utterance.

3.2.2. NER bias words list

As a second alternative, we use an NER model², to generate the complete word set for LibriSpeech. This approach is closer to real-world scenarios, where biasing lists for ASR systems often include contact names, phone numbers, personal names, and location names.

We split the words in the LibriSpeech text by spaces and apply the NER model to filter them one by one. After the completion of the filtering process, we successfully obtain a comprehensive bias words list consisting of 4365 words.

3.3. Training configurations

In our approach, we utilize edge-tts³ as the TTS model and Whisper-turbo⁴ as the ASR model. We employ a linear layer to project the 1280-dimensional features extracted from the Whisper-large-v3 model down to 368 dimensions. For the cross-attention mechanism, we utilize an 8-head configuration with a dropout rate of 0.1 during the training phase. The CNN consists of four layers. The output channels of these layers are set to 32, 64, 128, and 256 respectively, with a uniform kernel size of 3 for all layers. We apply an adaptive average pooling layer to aggregate the features into a single 256-dimensional vector.

During the training process, we adopt different learning rates for different bias words list. For the is12 deep bias words list, we set the learning rate to 0.00005. Given that the number

²<https://huggingface.co/FacebookAI/xlm-roberta-large-finetuned-conll103-english>

³<https://github.com/rany2/edge-tts>

⁴<https://huggingface.co/openai/whisper-large-v3-turbo>

of bias words in the NER bias words list in training is relatively small, we use a larger learning rate of 0.0001. Regarding the selection of negative words, 40% of the negative samples are chosen based on the edit distance, while the remaining 60% are randomly selected. This mixed-rate approach enriches the training data and improves the model’s robustness.

3.4. Evaluation metrics

We adopted two distinct metrics for evaluation, following [34]. First is the overall word error rate (WER), which gauges the word error rate across all words. Second is the B-WER, a biased metric that calculates the word error rate for words within the biasing list. The aim of contextualization is to boost the B-WER without notably increasing the WER.

4. Experimental Results

4.1. Results for Whisper with IS21 deep bias words list

4.1.1. Overall Evaluation

The WER and B-WER results for the LibriSpeech dataset, using the IS21 deep bias word list as prompts with the Whisper-turbo model, are summarized in Table 1. When the raw IS21 bias list (size $N = 100$) is used, the Whisper model achieves a relative reduction of approximately 25% in B-WER compared to the baseline model without prompts. However, when the bias list size increases to $N \geq 500$, the reduction in B-WER becomes negligible. In contrast, by selecting the top $k = 50$ bias words with the highest scores using our proposed scorer network, we observe a significant 40% relative reduction in B-WER. Notably, even with a smaller bias word list ($N = 100$), the reduction in B-WER remains above 30%.

4.1.2. Ablation studies on Prompt-based Biasing

To further validate the effectiveness of the proposed approach, we conducted experiments using different styles of prompts, as described in Section 2.1. The results, presented in Table 2, show that when only the true bias words are included, the B-WER achieves its optimal value. However, this improvement comes at the cost of a deterioration in WER, due to the hallucination phenomenon. When the number of bias words is too large, the additional content introduced by the spoken-style prompt distracts the model’s attention from the true bias words, leading to increased errors in the transcriptions.

We also investigated the effect of word order on contextual ASR performance. For the IS21 deep bias word list, in the random order configuration, true bias words are randomly interspersed with distractors. We also tested two additional ordered arrangements. In the ascending score order, bias words with higher probabilities are placed at the end of the list, while in the descending score order, words with higher likelihoods are positioned at the beginning. All scores were generated using our proposed scorer network. As shown in Table 3, when prompts are directly input to the Whisper model, the model performs best when the true bias words are placed at the end of the list, i.e., following the ascending order.

4.2. Generalization of our method

To test the generalization of our method, we apply it to different contextual ASR models and different bias word lists.

Generalization to different contextual ASR systems: To assess the generalization of our approach to other contextual

Table 1: WER (%) and B-WER (%) results for the Whisper model with prompts using the IS21 deep bias words list with the best in **bold** and the second underlined. The baseline is the Whisper-turbo model without any bias prompt. All the results are obtained using the naive prompt described in Section 2.1. When the proposed scorer network is used, we rank the key words in ascending order and select the top $k = 50$ key words with the highest scores; otherwise, we directly use the entire IS21 deep bias words list. The reported metrics follow the format: test-clean / test-other.

Method	N=100		N=500		N=1000		N=2000	
	WER	B-WER	WER	B-WER	WER	B-WER	WER	B-WER
DB-RNNT [34]	2.8 / 8.1	7.4 / 17.7	2.9 / 8.3	8.1 / 19.1	3.0 / 8.5	8.5 / 20.5	3.0 / 8.8	8.9 / 21.8
DB-RNNT+NNLM [34]	2.0 / 5.9	<u>5.7</u> / 14.1	2.1 / 6.1	<u>6.2</u> / 15.1	2.1 / 6.4	<u>6.7</u> / 17.2	2.3 / 6.6	<u>7.3</u> / 18.9
BPB [35]	2.8 / 5.6	6.0 / <u>12.0</u>	3.2 / 6.3	7.0 / <u>13.5</u>	3.5 / 7.3	7.7 / <u>15.8</u>	-	-
Whisper-turbo [36]	3.2 / 5.3	10.4 / 19.7	3.2 / 5.3	10.4 / 19.7	3.2 / 5.3	10.4 / 19.7	3.2 / 5.3	10.4 / 19.7
+ prompt	2.7 / 4.8	7.8 / 15.2	2.9 / <u>5.2</u>	9.5 / 19.1	2.9 / <u>5.2</u>	10.0 / 19.3	2.9 / <u>5.2</u>	9.8 / 19.5
+ + proposed	<u>2.3</u> / 4.2	5.4 / 10.0	<u>2.4</u> / 4.3	5.4 / 10.7	<u>2.4</u> / 4.4	5.6 / 11.3	<u>2.5</u> / 4.5	5.8 / 11.7
+ TCPGen [37]	3.8 / 6.5	11.0 / 19.3	3.6 / 6.6	10.8 / 19.7	3.9 / 6.9	11.4 / 20.4	4.2 / 7.0	12.0 / 21.1
+ + proposed	3.9 / 6.2	9.5 / 16.8	3.8 / 6.2	9.5 / 17.0	3.8 / 6.5	9.8 / 17.2	4.0 / 6.4	10.2 / <u>18.2</u>

Table 2: The influence of different prompt types, with the best result in **bold** and the second best underlined. The baseline is the Whisper-turbo without bias prompts, and the "no distractor" scenario uses only the bias words in the reference text. Other configurations use the IS21 deep bias list with $N = 100$.

prompt type	WER		B-WER	
	test-clean / test-other	test-clean / test-other	test-clean / test-other	test-clean / test-other
baseline	3.23 / 5.33		10.38 / 19.72	
no distractor	4.93 / 7.64		5.59 / 10.02	
naive	2.67 / 4.78		<u>7.78</u> / <u>15.21</u>	
spoken	<u>2.91</u> / <u>4.95</u>		8.30 / 15.50	

Table 3: Table showing the impact of bias word order on model performance, with the best WER and B-WER in **bold**. All bias word lists are based on the IS21 list, with each sentence having $N = 100$ bias words.

order	WER		B-WER	
	test-clean / test-other	test-clean / test-other	test-clean / test-other	test-clean / test-other
random	2.67 / 4.78		7.78 / 15.21	
ascending	2.37 / 4.14		5.73 / 10.15	
descending	2.97 / 5.19		10.03 / 19.44	

ASR methods, we also conducted experiments by feeding the biasing words into the TCPGen-based Biasing Whisper [37] as shown in Table 1. TCPGen is a tree-constrained pointer generator model that provides features to Whisper. Compared with the prompt-based method, TCPGen offers greater robustness when handling a large number of biasing words. However, despite its robustness, as the number of biasing words increases, the performance still experiences a relative decline of about 10%. By using the scorer proposed in our study for ranking and selection, the performance can be relatively improved by about 13% even when there are a large number of biasing words, reaching a level comparable to that when there are fewer biasing words.

However, it is still important to note that the word order remains a significant factor for TCPGen. In the study by [37], distractors are directly appended to the end of the bias list, which represents the ground-truth order. After employing the proposed scorer network for ranking and selection in the de-

scending order, the system performance can approach that of the ground-truth order.

Generalization of Different bias list: To better approximate real-world scenarios, we also examine the results using 4365 named-entity words filtered by the NER model from the LibriSpeech-clean-100 dataset. The results are presented in Table 4. For the Whisper-turbo model with prompts, selecting the top $k = 50$ words results in a significant relative reduction of B-WER (30%).

Table 4: Results for the Whisper model with prompt using NER bias words list of LibriSpeech with the best in **bold**. We use scorer network to rank and select top $k = \{10, 50, 100, 200\}$ words with the highest score from the all 4365 NER words list.

top k	WER		B-WER	
	test-clean / test-other	test-clean / test-other	test-clean / test-other	test-clean / test-other
0 (baseline)	3.23 / 5.33		16.70 / 25.15	
10	2.93 / 5.23		13.51 / 19.46	
50	2.78 / 5.16		11.51 / 17.74	
100	2.86 / 5.17		12.92 / 18.64	
200	2.85 / 5.13		12.92 / 18.64	

5. Conclusion

In this work, we try to address the challenge of contextual ASR models in handling numerous bias words. The proposed scorer network for bias word ranking and selection proves effective, achieving over 40% relative reduction in B-WER on the LibriSpeech dataset with the IS21 bias word list. It also shows good generalization across different models and different bias word lists. Insights into prompt types and word orders are provided, and this research paves the way for future progress in ASR systems, enabling better handling of rare and domain-specific words in real-world applications.

6. 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

7. References

- [1] J. Li *et al.*, “Recent advances in end-to-end automatic speech recognition,” *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.
- [2] 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.
- [3] 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.
- [4] K. B. Hall, E. Cho, C. Allauzen, F. Beaufays, N. Coccaro, K. Nakajima, M. Riley, B. Roark, D. Rybach, and L. Zhang, “Composition-based on-the-fly rescoring for salient n-gram biasing,” in *Interspeech*, 2015.
- [5] I. Williams, A. Kannan, P. S. Aleksic, D. Rybach, and T. N. Sainath, “Contextual speech recognition in end-to-end neural network systems using beam search,” in *Interspeech*, 2018.
- [6] Y. Qian, X. Gong, and H. Huang, “Layer-wise fast adaptation for end-to-end multi-accent speech recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 30, 2022.
- [7] D. Zhao, T. N. Sainath, D. Rybach, P. Rondon, D. Bhatia, B. Li, and R. Pang, “Shallow-fusion end-to-end contextual biasing,” in *Interspeech*, 2019.
- [8] Z. Chen, M. Jain, Y. Wang, M. L. Seltzer, and C. Fuegen, “End-to-end contextual speech recognition using class language models and a token passing decoder,” *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6186–6190, 2018.
- [9] R. Huang, O. Abdel-Hamid, X. Li, and G. Evermann, “Class LM and word mapping for contextual biasing in end-to-end ASR,” in *Interspeech*, 2020.
- [10] D. Le, G. Keren, J. Chan, J. Mahadeokar, C. Fuegen, and M. L. Seltzer, “Deep shallow fusion for RNN-T personalization,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 251–257.
- [11] 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.
- [12] X. Shi, Y. Yang, Z. Li, and S. Zhang, “SeACo-Paraformer: A non-autoregressive ASR system with flexible and effective hot-word customization ability,” *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10 346–10 350, 2023.
- [13] X. Gong, Y. Wu, J. Li, S. Liu, R. Zhao, X. Chen, and Y. Qian, “Longfmt: Long-form speech recognition with factorized neural transducer,” in *ICASSP 2023*. IEEE, 2023, pp. 1–5.
- [14] G. Pundak, T. N. Sainath, R. Prabhavalkar, A. Kannan, and D. Zhao, “Deep context: End-to-end contextual speech recognition,” *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 418–425, 2018.
- [15] M. Jain, G. Keren, J. Mahadeokar, and Y. Saraf, “Contextual RNN-T for open domain ASR,” in *Interspeech*, 2020.
- [16] K. Huang, A. Zhang, Z. Yang, P. Guo, B. Mu, T. Xu, and L. Xie, “Contextualized end-to-end speech recognition with contextual phrase prediction network,” in *Interspeech*, 2023.
- [17] M. Han, L. Dong, S. Zhou, and B. Xu, “Cif-based collaborative decoding for end-to-end contextual speech recognition,” *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6528–6532, 2020.
- [18] 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*. IEEE, 2023, pp. 1–5.
- [19] Z. Yang, S. Sun, X. Wang, Y. Zhang, L. Ma, and L. Xie, “Two stage contextual word filtering for context bias in unified streaming and non-streaming transducer,” in *Interspeech*, 2023.
- [20] X. Yang, W. Kang, Z. Yao, Y. Yang, L. Guo, F. Kuang, L. Lin, and D. Povey, “PromptASR for contextualized ASR with controllable style,” in *ICASSP 2024*. IEEE, 2024, pp. 10 536–10 540.
- [21] H. Futami, E. Tsunoo, Y. Kashiwagi, H. Ogawa, S. Arora, and S. Watanabe, “Phoneme-aware encoding for prefix-tree-based contextual ASR,” in *ICASSP 2024*. IEEE, 2024, pp. 10 641–10 645.
- [22] 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 Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1803–1815, 2024.
- [23] M. Wang, W. Han, I. Shafran, Z. Wu, C.-C. Chiu, Y. Cao, Y. Wang, N. Chen, Y. Zhang, H. Soltau, P. K. Rubenstein, L. Zilka, D. Yu, Z. Meng, G. Pundak, N. Siddhartha, J. Schalkwyk, and Y. Wu, “SLM: Bridge the thin gap between speech and text foundation models,” *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8, 2023.
- [24] Z. Chen, H. Huang, A. Y. 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,” *ICASSP 2024*, pp. 13 521–13 525, 2023.
- [25] C. Sun, Z. Ahmed, Y. Ma, Z. Liu, Y. Pang, and O. Kalinli, “Contextual biasing of named-entities with large language models,” *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10 151–10 155, 2023.
- [26] X. Gong, A. Lv, Z. Wang, and Y. Qian, “Contextual biasing speech recognition in speech-enhanced large language model,” *Interspeech*, 2024.
- [27] G. Yang, Z. Ma, Z. Gao, S. Zhang, and X. Chen, “CTC-assisted LLM-based contextual ASR,” *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 126–131, 2024.
- [28] A. Andrusenko, A. Laptev, V. Bataev, V. Lavrukhin, and B. Ginsburg, “Fast context-biasing for CTC and transducer ASR models with CTC-based word spotter,” *Interspeech*, 2024.
- [29] 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.
- [30] 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.
- [31] Y. Li, M. Zhang, C. Su, Y. Li, X. Qiao, M. Ren, M. Ma, D. Wei, S. Tao, and H. Yang, “A multitask training approach to enhance whisper with open-vocabulary keyword spotting,” *Interspeech 2024*, 2023.
- [32] A. Navon, A. Shamsian, N. Glazer, G. Hetz, and J. Keshet, “Open-vocabulary keyword-spotting with adaptive instance normalization,” *ICASSP 2024*, pp. 11 656–11 660, 2023.
- [33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.
- [34] D. Le, M. Jain, G. Keren, S. Kim, Y. Shi, J. Mahadeokar, J. Chan, Y. Shangguan, C. Fuegen, O. Kalinli, Y. Saraf, and M. L. Seltzer, “Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion,” in *Interspeech*, 2021.
- [35] Y. Sudo, M. Shakeel, Y. Fukumoto, Y. Peng, and S. Watanabe, “Contextualized automatic speech recognition with attention-based bias phrase boosted beam search,” *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10 896–10 900, 2024.
- [36] 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.
- [37] G. Sun, X. Zheng, C. Zhang, and P. C. Woodland, “Can contextual biasing remain effective with whisper and gpt-2?” in *Interspeech*, 2023.