

# PROMPTASR FOR CONTEXTUALIZED ASR WITH CONTROLLABLE STYLE

Xiaoyu Yang, Wei Kang, Zengwei Yao, Yifan Yang, Liyong Guo, Fangjun Kuang, Long Lin, Daniel Povey

Xiaomi Corp. Beijing, China

{yangxiaoyu6, dpovey}@xiaomi.com

## ABSTRACT

Prompts are crucial to large language models as they provide context information such as topic or logical relationships. Inspired by this, we propose PromptASR, a framework that integrates prompts in end-to-end automatic speech recognition (E2E ASR) systems to achieve contextualized ASR with controllable style of transcriptions. Specifically, a dedicated text encoder encodes the text prompts and the encodings are injected into the speech encoder by cross-attending the features from two modalities. When using the ground truth text from preceding utterances as content prompt, the proposed system achieves 21.9% and 6.8% relative word error rate reductions on a book reading dataset and an in-house dataset compared to a baseline ASR system. The system can also take word-level biasing lists as prompt to improve recognition accuracy on rare words. An additional style prompt can be given to the text encoder and guide the ASR system to output different styles of transcriptions. The code is available at icefall<sup>1</sup>.

**Index Terms**— Contextualized ASR, Prompts, Transducer,

## 1. INTRODUCTION

External text information is commonly used to improve E2E ASR systems. Traditional approaches use external language models trained on the text corpora and re-rank the n-best hypotheses [1, 2] of ASR systems or perform shallow fusion [3] to modify the posterior predicted by the ASR system.

Recently, various methods have been proposed to utilize contextual information to improve the accuracy of speech recognition [4, 5, 6, 7], namely contextualized ASR. Depending on the form of the context, most existing contextualized ASR systems fall into two categories: word-level context and utterance-level context. Word-level context biasing aim to improve the recognition accuracy of rare words such as contact names or application names. Sun et.al [5] used a tree-constrained pointer generator to boost the posterior of words in a context list if the prefix matches during decoding. Huang et.al [6] proposed a neural network-based method to improve rare-word recognition on various E2E ASR architectures with a context phrase prediction network. Unlike word-level

context, utterance-level context carries more sophisticated information such as topic and logical relationships. Text embeddings encoded by BERT [8] are utilized [9] to improve ASR performance in multi-turn dialogues. Similarly, Chang et. al [7] improve long-form ASR [10] performance on neural transducer [11] using self-attentive embeddings from BERT. Li et.al [12] used LLaMa [13] as the decoder of a speech encoder to facilitate domain adaptation through text prompts such as titles and topic descriptions.

In large language models, prompts are crucial to the correctness, fluency and format of the generated text [14, 15]. Inspired by this, we propose a novel E2E ASR framework named PromptASR, which utilizes text prompts for contextualized speech recognition. In specific, a dedicated text encoder is added to the E2E ASR system to ingest two types of prompt: content prompt and style prompt. The prior provides contextual information and the latter specifies the style of desired ASR transcriptions (e.g. casing and punctuation). The encoded prompts are injected to the ASR encoder via cross-attention with hidden speech representations. Unlike most existing approaches for contextualized ASR that are specialized for either word or utterance-level context, PromptASR is able to benefit from both of them. When decoded with the ground truth preceding text as content prompt, PromptASR achieves 21.9% and 6.3% relative word-error-rate (WER) reduction compared to a baseline ASR system on a book reading dataset and an in-house dataset. On a word-level context biasing task [4], PromptASR achieves 13.4% relative WER reduction even with biasing lists containing 1000 distractors. Finally, we show that the style prompt effectively guides the style of transcriptions. It is noteworthy that Whisper [16] mentioned having a similar prompting mechanism, but the details are not included in their paper.

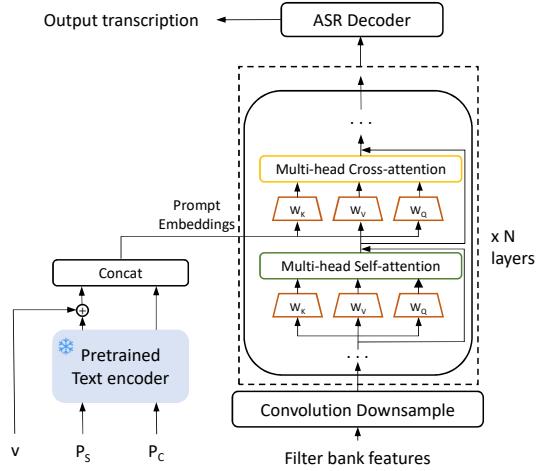
## 2. PROMPTASR

### 2.1. System Architecture

The architecture of PromptASR is illustrated in Fig. 1. It consists of a pre-trained text encoder  $Enc^T$ , a speech encoder  $Enc^A$  and an ASR decoder  $Dec^A$ . The text encoder  $Enc^T$  processes prompts and generates text embeddings.  $Enc^A$  consists of N transformer-like layers, each with a cross-attention module (with residual connection) placed after the self-attention module. Each layer receives not only the acoustic embed-

<sup>1</sup><https://github.com/k2-fsa/icefall>

dings, but the text embeddings encoded by  $Enc^T$ . The fusion between the text modality and speech modality is achieved by cross attention, where the text embeddings serve as query and acoustic hidden states serve as key. The whole system can be trained with any ASR objective functions.



**Fig. 1.** The architecture of PromptASR. The module in the dashed block is a transformer-like layer with cross-attention (other modules omitted). The text embeddings are injected as key/value pairs in the cross-attention module.

## 2.2. Prompts

Two types of prompts are defined in PromptASR: content prompt and style prompt. Content prompt should contain semantic and context-related information, which are usually in the form of sentences or a list of rare-words to be boosted. Most existing ASR systems produce normalized transcriptions, requiring inverse text normalization for production scenarios. Motivated by this, we would like the model to output different styles of transcription given different style prompts. The style prompt should indicate the desired style of the ASR transcription, such as casing and punctuation. During training, the style of the target text should always match the style prompt. Two training samples of PromptASR with different styles are shown in Table 1.

Style Prompt	WITHOUT CASING OR PUNCTUATION
Content Prompt	Welcome to the UEFA Champions League final!
Reference text	TODAY'S MATCH IS BETWEEN REAL MADRID AND LIVERPOOL
Style Prompt	Mixed-cased English with punctuation
Content Prompt	Welcome to the UEFA Champions League final!
Reference text	Today's match is between Real Madrid and Liverpool.

**Table 1.** Two training samples with different style prompts.

The tokenized content prompt  $\mathbf{P}_c = \mathbf{P}_{c,1}, \dots, \mathbf{P}_{c,n}$   $\mathbf{P}_c$  and style prompt  $\mathbf{P}_s = \mathbf{P}_{s,1}, \dots, \mathbf{P}_{s,m}$  are feed to the text en-

coder to produce prompt embeddings  $\mathcal{E}_c \in \mathbb{R}^{n \times c}$  and  $\mathcal{E}_s \in \mathbb{R}^{m \times c}$ . To distinguish between content prompts and style prompts, a trainable style indicator vector  $v \in \mathbb{R}^c$  is added to the embeddings of the style prompts. The forward process of PromptASR is formulated in Eqn 1.

$$\begin{aligned} \mathcal{E}_c &= Enc^T(\mathbf{P}_c); \\ \mathcal{E}_s &= Enc^T(\mathbf{P}_s) + v; \\ \mathcal{G} &= Enc^A(\text{Concat}(\mathcal{E}_c, \mathcal{E}_s), \mathbf{X}); \\ \mathbf{y} &= Dec^A(\mathcal{G}), \end{aligned} \quad (1)$$

where  $\mathbf{X}$  is the input speech features and  $\mathbf{y}$  is the output transcription. **Concat** is the operation of concatenating two tensors along time axis.

To deal with situations where prompts are unavailable, both prompts are dropped out by a small probability during training so that the model learns to transcribe without prompts. In real-life scenarios, the content prompts can be completely irrelevant to the current utterance, requiring the model to learn to ignore such prompts. Therefore, a small proportion of the content prompts within the mini-batch are exchanged to simulate this scenario for better robustness to irrelevant prompts.

## 3. EXPERIMENT SETUP

### 3.1. Dataset

The open-sourced dataset Libriheavy [17] is chosen as one of the training sets of PromptASR, as each utterance in Libriheavy is provided with a ground truth transcription and its preceding text of length 1000 bytes. Casing and punctuation of both texts are preserved. The medium subset containing around 5000 hours transcribed book readings is adopted in this work. The official Libriheavy test-clean and test-other sets are adopted for evaluation, which are approximately 20% harder (in terms of word error rates) than the LibriSpeech [18] test-clean and test-other sets. Additional 2000 hours recordings of conversations and podcasts covering different topics are also collected from the National Public Radio (NPR). An official transcript with casing and punctuation is provided for each recording. The text-audio alignments are obtained based on the Levenshtein distance between the output of a pre-trained ASR system and the official transcript and verified by human experts. The 1000-byte-long preceding text for each utterance is also extracted according to the alignment. 18 hours of recordings are hold-out and form the NPR evaluation set.

### 3.2. Model Selection

The pre-trained BERT [8] model is selected as the text encoder as it captures contextualized information through the masked language modeling pre-training. In addition, we also pre-train two transformers using the BERT objective on the text data of Libriheavy or NPR. Note that there are no overlaps between the training and testing books/recording. The parameters of the text encoder are frozen during training.

The ASR system is a neural transducer with a Zipformer [19] speech encoder. A stateless decoder [20] seeing two previous tokens and a joint network is added and pruned. RNNT [21] loss is used as the training objective. Chunk-wise streaming [19] is adopted for training streaming ASR models.

### 3.3. PromptASR Training

Each utterance’s preceding 1000 bytes are used as the content prompt during training. Two types of style transforms are predefined: **Upper-Cased** without punctuation (UC) and **Mixed-Cased with Punctuation** (MCP). A style is sampled for each utterance in the mini-batch and is applied to its style prompt and reference prompt. The **MCP** has a higher sampling probability (0.7) since it is more production-friendly. The style prompt is a sub-string of content prompt from other samples in the same mini-batch. To construct word-list based content prompt, the number of total appearances of each word in the training set is counted and the words outside the most common 10000 words are regarded as rare words. A context list is formed for each training sample by picking up rare words present in this sample and adding 50-100 randomly distractors. Together with the preceding text, both types of content prompts are used by probability of 0.5.

During training, SpecAug [22] and MUSAN [23] are adopted to augment the training data. Speed perturbation is not used. The 80-D mel filter bank features are used as the input acoustic features. Byte-pair encoding [24] with byte fallback is used as modelling units and the vocabulary size is 500. The model is trained for 50 epochs on Libriheavy medium subset, and 60 epochs on the NPR dataset. The checkpoints of last ten epochs are averaged and beam search of size 4 is used for inference.

### 3.4. Metrics

Two types of evaluation scenarios are investigated. First, each test sample is given its ground truth preceding 1000 bytes as content prompt during decoding. This can be seen as the performance upper bound of the PromptASR model when dealing with utterance-level content prompt. However, the ground truth preceding text is not always available in real-life scenarios and the model has to rely on the erroneous decoding results of the previous utterances. Therefore, a 15-hour long-form recordings test set is also collected from NPR to validate the long-form ASR performance. The long recordings are split into individual sentences without overlap. The average length of the recording is 20 minutes. The decoding results from previous sentences are concatenated to construct the content prompt for the current utterance and a fixed style prompt irrelevant to the recording is used.

To test the capability of PromptASR on word-level context biasing, the biasing list for LibriSpeech [18] test sets in [4] is used. Each utterance in the test set is provided with a biasing word list containing biasing words and distractors. Note that the biasing lists of some utterances are purely distractors.

## 4. EXPERIMENT RESULTS

### 4.1. Utterance-level Context Biasing

Experiments are first carried out to validate the benefit of content prompts, where the ground truth preceding text is used as content prompt during decoding. Baseline neural transducer models without text encoder are trained with UC and MCP styles separately. The word-error-rates (WERs) are shown in Table 2 and the following points can be drawn. First, PromptASR model significantly improves the WERs owing to the contextual information from the content prompts. For non-streaming models, relative WER reductions (WERRs) of 21.9% and 6.8% are achieved compared to the baseline (**B1-UC**) on the Libriheavy test-other and NPR with the in-domain text encoder using style UC. Similar relative WERRs of 20.0% and 7.8% are observed for streaming models. If no content prompt is given, PromptASR still achieves comparable results as the baseline model, suggesting that the model is robust to decode without any prompts. Second, pre-training the text encoder on in-domain data further improves the performance on the in-domain test sets. Finally, the style change in PromptASR does not affect the WER. After normalizing the transcript with UC style, decoding with MCP style yields similar WERs as using UC style.

Model	Content Prompt	Style Prompt	Libriheavy		NPR
			clean	other	
<b>Non-streaming Models</b>					
<b>B1-UC</b>	-	-	3.0	6.72	5.09
<b>B1-MCP</b>	-	-	3.11 (10.4)	6.74 (14.3)	5.26 (9.5)
BERT [8]	-	-	3.14	6.71	5.26
Text Enc	✓	UC	2.82	6.03	4.89
	✓	MCP	2.64 (9.3)	5.55 (12.5)	4.95 (8.48)
In-domain	-	-	3.09	6.8	5.28
Text Enc	✓	UC	2.48	5.25	4.74
	✓	MCP	2.38 (9.0)	5.03 (12.2)	4.8 (8.4)
<b>Streaming Models</b>					
<b>B2-UC</b>	-	-	3.73	8.19	6.77
In-domain	-	-	3.7	8.13	6.96
Text Enc	✓	UC	3.01	6.55	6.24
	✓	MCP	2.93 (10.5)	6.29 (14.4)	6.29 (10.8)

**Table 2.** WERs (%) of baseline models and PromptASR models on different test sets. For MCP style, WERs before (in brackets) and after UC style text normalization are reported.

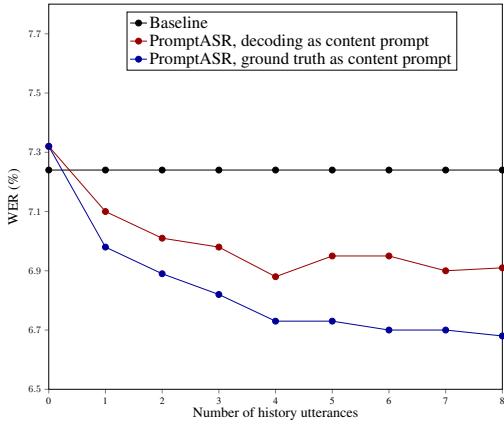
### 4.2. Long-form ASR

The performance of PromptASR on long-form ASR is investigated and results are shown in Fig 2. The non-streaming PromptASR model with in-domain text encoder trained on NPR is decoded in UC style with either erroneous decoding results (red) or the ground truth transcripts (blue) of the history utterances as content prompt. For reference, the WER of a baseline ASR system **B1-UC** from Table 2 is also plotted (black). Though both content prompts reduce the WER compared to the baseline, the gain from using erroneous decoding

Model	Word-level LibriSpeech Biasing										Utterance-level Libriheavy Biasing	
	No Biassing		N=0		N=100		N=500		N=1000			
	clean	other	clean	other	clean	other	clean	other	clean	other	clean	other
<b>B1-UC</b>	2.46	5.11	2.46	5.11	2.46	5.11	2.46	5.11	2.46	5.11	3.0	6.72
<b>M1</b>	2.45	5.09	2.37	4.90	2.49	5.36	2.62	5.71	2.69	5.82	2.48	5.25
<b>M2</b>	2.43	5.07	<b>1.73</b>	<b>4.0</b>	<b>1.73</b>	<b>4.07</b>	<b>2.0</b>	<b>4.45</b>	<b>2.13</b>	<b>4.67</b>	2.59	5.55

**Table 3.** WERs (%) with different context list size on the LibriSpeech biasing task.  $N$  is the number of distractors added to the context list. The WERs of the Libriheavy utterance-level biasing test are also shown.

results of preceding utterances is smaller and it fails to further improve the WER with a history of longer than 4 utterances. This could be caused by the wrong prediction of named entities or keywords, which undermines the benefit from context information for PromptASR.



**Fig. 2.** WERs (%) of baseline model and PromptASR models after normalization on the long-form ASR task.

### 4.3. Word-level Context Biasing

The potential of applying PromptASR for rare-word recognition is investigated. All the models in Table 3 are trained on Libriheavy medium subset. Note that only the in-domain text encoder can be used for this task, as the official release of BERT has a maximum input length constraint of 512 tokens. **B1-UC** is the non-streaming baseline model from Table 2. **M1** and **M2** are PromptASR models sharing the same text encoder pre-trained on the Libriheavy large subset (disjoint from LibriSpeech test sets). **M1** only uses the previous utterances as content prompt during training, while **M2** additionally uses content prompts constructed from rare-words list as described in Sec 3.3. The WERs with different sizes of context lists on the LibriSpeech biasing task are shown in Table 3. All three models yield similar performance without biasing lists. When decoded with biasing lists, **M1** fails to benefit from the word-level context. However, only at the cost of slight WER degradation on the utterance-level context biasing (last two columns), constructing content prompts from rare-words list (**M2**) during training significantly helps word-level contextual biasing, achieving relative WERRs of 29.7% and 21.7% at  $N = 100$  compared to the baseline model. However, as the number of distractors increases, the performance gain

decreases very quickly and the relative WERRs at  $N = 1000$  are 13.4% and 8.6%. One possible reason is that the context list used during training is shorter than 100 words, thus the model did not generalize well to larger  $N$ . Despite this, the relative improvements are still comparable with the existing neural network-based context biasing methods [6], which yields 10.6% and 12.6% relative WERRs with  $N = 1000$ .

### 4.4. Output Examples

A few examples of the output of PromptASR model is shown below. In Table 4, the first block shows an example where the model outputs transcriptions with accurate casing and punctuations using MCP style. The second block shows that the PromptASR model corrects the ASR output with the help of content prompt – it learns from the content prompt that the topic is about “horse”, and guides the model to output the in-domain word “phaeton” instead of a made-up word “faithon”.

style prompt	Mixed-cased English with punctuations.
PromptASR output	“Do you believe in some education?” asked Mary Taylor.
content prompt	... I knew how hard it was upon slow-paced <b>horses</b> to be put with fast ones; ...
Baseline output	She was often used in the <b>faithon</b> , and was very much liked by some of the ladies.
PromptASR output	She was often used in the <b>phaeton</b> , and was very much liked by some of the ladies.

**Table 4.** Outputs of PromptASR versus normal ASR system.

## 5. CONCLUSION

We propose the PromptASR framework, which performs contextualized ASR with controllable style of transcriptions. By passing either the ground truth transcript or decoded transcript of previous utterances as content prompt, the PromptASR model utilizes the cross-utterance context and improves the WER compared to a baseline ASR system. If the content prompt is a list of biasing words, the PromptASR model can also perform word-level biasing and achieve significant WER reduction on biasing words while being robust to distractors. The model can switch the output style (e.g. casing and punctuation) given different style prompts. In the future, we hope to explore more efficient utilization of text embeddings to reduce the computational cost. We are also very interested in how to incorporate large language models into PromptASR.

## 6. REFERENCES

- [1] Richard Schwartz and Steve Austin, “A comparison of several approximate algorithms for finding multiple ( $N$ -best) sentence hypotheses,” in *Proc. ICASSP*, Toronto, 1991.
- [2] Xianrui Zheng, Chao Zhang, and Philip C Woodland, “Adapting GPT, GPT-2 and BERT language models for speech recognition,” in *Proc. ASRU*, Cartagena, 2021.
- [3] Anjuli Kannan, Yonghui Wu, Patrick Nguyen, Tara N Sainath, Zhijeng Chen, and Rohit Prabhavalkar, “An analysis of incorporating an external language model into a sequence-to-sequence model,” in *Proc. ICASSP*, Calgary, 2018.
- [4] Duc Le, Mahaveer Jain, Gil Keren, Suyoun Kim, Yangyang Shi, Jay Mahadeokar, Julian Chan, Yuan Shangguan, et al., “Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion,” in *Proc. Interspeech*, Brno, 2021.
- [5] Guangzhi Sun, Chao Zhang, and Philip C Woodland, “Tree-constrained pointer generator for end-to-end contextual speech recognition,” in *Proc. ASRU*, Cartagena, 2021.
- [6] Kaixun Huang, Ao Zhang, Zhanheng Yang, Pengcheng Guo, et al., “Contextualized end-to-end speech recognition with contextual phrase prediction network,” in *Proc. Interspeech*, Dublin, 2023.
- [7] Shuo-Yiin Chang, Chao Zhang, Tara N Sainath, Bo Li, and Trevor Strohman, “Context-aware end-to-end ASR using self-attentive embedding and tensor fusion,” in *Proc. ICASSP*, Rhodes, 2023.
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. NAACL*, Minneapolis, 2019.
- [9] Kai Wei, Thanh Tran, Feng-Ju Chang, et al., “Attentive contextual carryover for multi-turn end-to-end spoken language understanding,” in *Proc. ASRU*, Cartagena, 2021.
- [10] Arun Narayanan, Rohit Prabhavalkar, Chung-Cheng Chiu, David Rybach, Tara N Sainath, and Trevor Strohman, “Recognizing long-form speech using streaming end-to-end models,” in *Proc. ASRU*, 2019.
- [11] A. Mohamed A. Graves and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *Proc. ICASSP*, Vancouver, 2013.
- [12] Yuang Li, Yu Wu, Jinyu Li, and Shujie Liu, “Prompting large language models for zero-shot domain adaptation in speech recognition,” *arXiv preprint arXiv:2306.16007*, 2023.
- [13] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, et al., “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.
- [14] Brian Lester, Rami Al-Rfou, and Noah Constant, “The power of scale for parameter-efficient prompt tuning,” in *Proc. EMNLP*, Punta Cana, 2021.
- [15] Laria Reynolds and Kyle McDonell, “Prompt programming for large language models: Beyond the few-shot paradigm,” in *Proc. CHI*, New York, 2021.
- [16] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavy, and Ilya Sutskever, “Robust speech recognition via large-scale weak supervision,” in *Proc. ICML*, Hawaii, 2023.
- [17] Wei Kang et al., , ” [https://github.com/k2-fsa/text\\_search/tree/master/examples/libriheavy](https://github.com/k2-fsa/text_search/tree/master/examples/libriheavy), 2023.
- [18] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: an ASR corpus based on public domain audio books,” in *Proc. ICASSP*, Brisbane, 2015.
- [19] D. Povey, , ” <https://github.com/k2-fsa/icefall/tree/master/egs/librispeech/ASR/zipformer/zipformer.py>, 2023.
- [20] Mohammadreza Ghodsi, Xiaofeng Liu, James Apfel, Rodrigo Cabrera, and Eugene Weinstein, “RNN-transducer with stateless prediction network,” in *Proc. ICASSP*, Barcelona, 2020.
- [21] Fangjun Kuang, Liyong Guo, Wei Kang, Long Lin, Mingshuang Luo, Zengwei Yao, and Daniel Povey, “Pruned RNN-T for fast, memory-efficient ASR training,” in *Proc. Interspeech*, Incheon, 2022.
- [22] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, et al., “SpecAugment: A simple data augmentation method for automatic speech recognition,” in *Proc. Interspeech*, Graz, 2019.
- [23] David Snyder, Guoguo Chen, and Daniel Povey, “MUSAN: A Music, Speech, and Noise Corpus,” *arXiv:1510.08484v1*, 2015.
- [24] Rico Sennrich, Barry Haddow, and Alexandra Birch, “Neural machine translation of rare words with subword units,” in *Proc. ACL*, Berlin, 2016.